DESIGN CHALLENGES AND CAD SOLUTIONS FOR LOW POWER AND RELIABLE MONOLITHIC 3D ICS

A Dissertation
Presented to
The Academic Faculty

by

Sandeep Kumar Samal

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2017

DESIGN CHALLENGES AND CAD SOLUTIONS FOR LOW POWER AND RELIABLE MONOLITHIC 3D ICS

Approved by:

Dr. Sung Kyu Lim, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Azad J. Naeemi
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Hyesoon Kim
College of Computing
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: February 23, 2017

Dedicated to my family:

my parents Mrs. Sandhya Samal and Dr. Parshuram Samal

and my sister Dr. Priyanka Samal

for their endless love, support and encouragement.....

ACKNOWLEDGEMENTS

The successful journey of achieving the doctoral degree has been a great experience. It involved the guidance, help and support of many individuals, both in academic and social aspects of life. I would like to take this opportunity to express my deepest gratitude to all those who have helped me during the course of my study.

First of all, I would like to give my heartfelt thanks to Dr. Sung Kyu Lim, my PhD advisor, for his exceptional research guidance, advice and academic support. He provided me with this great opportunity to pursue the highest academic degree in one of the world's best academic and research institutions in our field. His research experience, ideas and inputs, both in my research projects and in other aspects of professional development, have helped me significantly during my PhD

I would like to thank Dr. Arijit Raychowdhury and Dr. Saibal Mukhopadhyay for providing valuable feedback on my research. I would also like to thank Dr. Azaad Naeemi and Dr. Hyesoon Kim for their inputs and suggestions while serving on my dissertation defense committee.

My graduate student life at Georgia Tech has been a great experience in terms of research and academic learning. I had the opportunity to study different courses and carry out research on very interesting topics. While the major credit for this goes to my PhD advisor, I also thank the faculty here at Georgia Tech, whose classes helped me better understand the various topics in electrical and computer engineering.

My research work was in collaboration with multiple sponsors and organizations, primarily Qualcomm Inc. and Globalfoundries Inc. I would like to thank Dr. Kambiz Samadi, Pratyush Kamal and Dr. Yang Du from Qualcomm for their inputs and feedback on my

research. The Technology Research group in Globalfoundries provided me with a one-year industry internship experience directly related to my research. I achieved significant research success while working with them and am thankful to Dr. Deepak Nayak, Dr. Srinivasa Banna and Dr. Motoi Ichihashi for their help and support.

I would like to thank the past and current members of the GTCAD lab at Georgia Tech: Dr. Krit Athikulwongse, Dr. Xin Zhao, Dr. Daniel Limbrick, Dr. Young-Joon Lee, Dr. Moongon Jung, Dr. Shreepad Panth, Dr. Taigon Song, Dr. Yarui Peng, Mohit Pathak, Neela Lohith, Kyungwook Chang, Bon Woong Ku, and Kartik Acharya for their help, advice, and the brainstorming sessions and insightful discussions on research ideas. I also thank David Webb, Keith May, Peter Huynh, and Pamela Halverson in the School of ECE, for their technical and administrative support during the course of my study.

I would like to express my deepest gratitude to my parents Mrs. Sandhya Samal and Dr. Parshuram Samal, and my sister Dr. Priyanka Samal. This phase of my life would not have been possible without their encouragement, unconditional love, constant support, and patience.

Finally, I am thankful to all my friends in Atlanta who made me feel like home away from home. I made many new friends from different parts of the world, and got to learn so many new things. The good times spent with them have had a very positive influence on me, and helped me grow as a person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Monolithic 3D IC (M3D) is enabled by sequential vertical integration of extremely thin device layers with very high alignment precision. Unlike TSVs, monolithic inter-tier vias (MIVs) are miniscule ($<$100nm diameter) and can be used in large numbers within the design. MIVs are similar to inter metal layer vias and have negligible capacitance ($\ll$1fF) compared to TSVs. This helps in high integration density allowing numerous 3D connections which in turn reduce wirelength, reduce power and help in improving performance. However, as with any new technology, sequential integration comes with new challenges, both in design and fabrication.

The objective of this research is to study and quantify the major challenges in low power and reliable monolithic 3D IC design and to develop CAD solutions to address these challenges, in order to obtain maximum benefits from emerging monolithic 3D IC technology. The key design challenges in monolithic 3D ICs covered in this research include thermal modeling and optimization, 3D PDN design and analysis and addressing inter-tier performance difference and inter-tier interconnect difference in monolithic 3D ICs. In addition, near-threshold voltage 3D IC design and analysis is also presented.

The increased power density due to multiple device layers and the absence of bulk silicon substrate in the sequential layers complicates thermal management of monolithic 3D ICs. The major bottleneck of considering thermal aspect within the physical design process is the huge runtime required for accurate temperature analysis. Therefore, importance of thermal-aware design methodologies become more critical in 3D ICs and faster analysis techniques are necessary. In the first research work, a fast-accurate regression-based thermal modeling technique is developed for monolithic 3D ICs. Next, this model is incorporated into a monolithic 3D IC floorplanner to make it thermal-aware. Thermal modeling and floorplanning is carried out for both conventional packages with heat sink and modern

mobile packaging structure without any dedicated heat sink.

Multiple device layers also increase the current demand per unit area. This necessitates better power delivery network design techniques. However, power delivery network in top metal layers of BEOL stack of bottom device tiers reduce the routing resources for 3D signal wires, resulting in increased congestion and wirelength, which in turn increases switching power. This problem is severe in monolithic 3D ICs due to the presence of tens of thousands of MIVs which use top metal layers of bottom tiers. A detailed study is conducted to quantify this trade-off between signal and power routing unique to monolithic 3D ICs. Design optimization techniques are developed to minimize switching power overhead while satisfying power delivery constraints.

Due to the presence of FEOL and BEOL in the first tier of monolithic 3D ICs, the thermal budget of fabricating the sequential device layers is constrained. As a consequence, matching the device performance to that of a regular CMOS process becomes a major challenge. Also, tungsten is used for BEOL in bottom-tier, since copper cannot withstand higher temperatures. The system level impact of low performance transistors in the top-tier of monolithic 3D ICs is quantified and new tier-aware gate-level monolithic 3D IC design flow is presented to enable low power designs under practical settings. In the next research work, the adverse impact of of BEOL in the bottom-tier of monolithic 3D ICs is studied. New tier partitioning strategies are presented to mitigate performance degradation due to tungsten and to reduce congestion and metal layer usage in the bottom tier.

Lastly, a power-performance-cost analysis of monolithic 3D ICs is presented using realistic cost data. Monolithic 3D IC benefits are also compared across different technology nodes. Overall impact of monolithic 3D IC technology is compared with 2D ICs and TSV-based 3D ICs and design and performance targets are estimated.

CHAPTER I

INTRODUCTION

Device and technology scaling is becoming more and more challenging with reduced return on investments. Advanced lithography, multiple-patterning, interconnect material properties, variations etc. are major issues in conventional technology scaling below 45nm node. Three dimensional integrated circuit (3D IC) technology has been studied as one possible path to continue scaling benefits by placing devices on multiple layers. 3D ICs offer lower power, better performance and reduced footprint requirement compared to 2D ICs.

Majority of the research on 3D ICs is focused on Through Silicon Via (TSV) based 3D IC technology. TSVs enable the vertical integration of separate dies to form a single 3D chip. However, TSVs consume a lot of area, have a large capacitance, and have very large pitch requirements during fabrication. In addition, the state-of-the art TSV size is orders of magnitude larger than the device sizes at advanced nodes. This puts a restriction on the number of TSVs and the type of circuits that can be used, limiting them to mostly memory-on-logic designs or interposer based 2.5D technology [19, 56]. Therefore, the greater benefits of 3D IC are masked by these unfavorable characteristics of TSVs.

Monolithic 3D ICs (M3D) is another 3D IC technology enabled by the sequential integration of device layers in the vertical direction. M3D integration uses nano-scale monolithic inter-tier vias (MIVs) to connect the vertical device layers. MIVs are similar to regular metal-layer vias and their capacitance and area values are negligible compared to those of TSVs that are micro-scale. This allows the use of many such MIVs for vertical connections which enable significantly higher integration density than that of TSV-based 3D ICs. MIVs enable various design styles starting from coarse block-level 3D partitioning to very fine-grained 3D partitioning in both the gate level as well as intra-gate level i.e. transistor

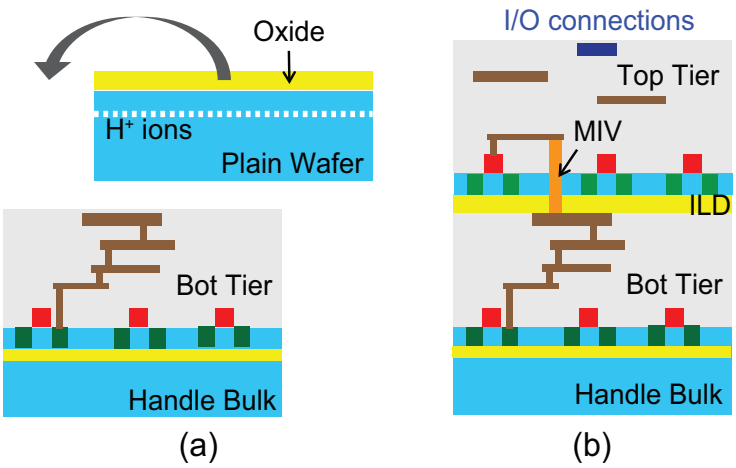


Figure 1: Sequential integration in monolithic 3D ICs.

level [45]. Though, the overall benefits of M3D technology is significant from a scaling perspective, the technology development is still under active research and there are many key challenges that need to be addressed before large-scale production.

This following sections gives an overview of the major design challenges in monolithic 3D ICs followed by the contributions and the organization of rest of this dissertation.

1.1 Challenges in Monolithic 3D ICs

The fabrication of high quality crystalline silicon on the top tier of already existing device layer was proposed by Batude *et al.* [5]. Figure 1 shows this process of sequential integration. First, the bottom-tier is fabricated with conventional 2D IC process resulting in regular quality transistors. An empty wafer with H^+ ions implanted below the silicon surface, and with thermal oxide grown over it, is bonded on to the first tier using low temperature molecular bonding (Figure 1a). The empty silicon wafer is then sheared along the H^+ line and polished. This is followed by the fabrication of monolithic inter-tier vias (MIVs), new transistors and interconnect layers for the second tier (Figure 1b).

As with any 3D IC technology, the multiple device layers in 3D ICs increase the power density of the chip and also the current demand per unit area. This in turn complicates the cooling issues and the power delivery network design. Even if power reduction is

achieved by going 3D, the increased power density affects the temperature, especially in the layers away from the heat sink or other equivalent cooling features in modern miniaturized electronics. Monolithic 3D IC is a fairly new technology and its thermal characteristics have not been studied earlier. The importance of thermal-aware design methodologies become more critical in monolithic 3D ICs and faster analysis techniques are necessary.

The presence of power delivery network (PDN) also reduces the routing resources and adversely affects the total signal wirelength. This impact is more severe in M3D than in 2D, due to heavy usage of top metals for 3D routing along with PDN routing. PDN is always important in any design, but due to sequential 3D layers with very high integration density in M3D, optimal PDN planning becomes much more critical. Therefore, with very high degree of integration density in M3D, it becomes more important to have a robust thermal-aware design scheme and power delivery optimization without affecting other design metrics.

In addition, due to the presence of FEOL and BEOL in the first tier, the thermal budget of fabricating the sequential device layers is constrained. As a consequence, matching the device performance to that of a regular CMOS process becomes a major challenge. Low thermal budget of sequential device layers mainly affects the dopant activation process, leading to reduced mobility. Researchers have tried Solid Phase Epitaxy at 625^{o}C [4] and laser annealing with low in-depth thermal diffusion [49] for dopant activation. However, there is a performance reduction in the resulting transistors at lower thermal budget. While new research progress has demonstrated that performance of devices processed at low temperature can potentially match performance of regular high temperature process devices [3], tungsten is used for BEOL in bottom-tier, since copper cannot withstand higher temperatures. With higher resistivity of tungsten, the quality of design gets affected. This necessitates the requirement of new design methods and CAD techniques to handle these practical issues and evaluate M3D technology.

Along with the required development of CAD tools and design infrastructure, another

key controlling factor in monolithic 3D ICs is the related cost and cycle-time overhead because of the presence of additional masks per wafer. The die size reduces in monolithic 3D ICs, improving the yield and the number of dies per wafer. However, additional number of layer can exceed the advantage of smaller die size. Therefore, cost reduction techniques, and power-performance-cost evaluation is an important research topic which has not been explored in prior works on monolithic 3D ICs.

1.2 Contributions

This dissertation focuses on understanding the details and impact of the challenges discussed above, and developing CAD solutions to address them efficiently. This work consists of five major research topics that can be summarized as follows.

- **Fast accurate thermal modeling and optimization for monolithic 3D ICs:** Monolithic 3D ICs can overcome the shortcomings of TSV-based 3D ICs. However, one major concern with 3D ICs in general is the increase in power density which leads to high temperature values and thermal issues. Thermal-aware monolithic 3D IC design is necessary in order to justify their overall advantages over 2D ICs and over TSV-based 3D ICs. In this dissertation, a fast and accurate non-linear regression based temperature evaluation model is developed after detailed study and analysis of the thermal characteristics of M3D and its comparison with thermal properties of TSV-based 3D ICs. The model is package-aware, and can handle both conventional packages with heat sink and modern mobile package without any heat sink. Next, the developed model is incorporated into a thermal-aware 3D floorplanner for temperature-optimized 3D IC floorplans. Thermal impact study of material properties and thickness values is also carried out for package optimization

- **Power delivery network (PDN) impact and optimization in full-chip monolithic 3D ICs:** PDN worsens routing congestion more severely in monolithic 3D ICs than in 2D designs due to the significant reduction in resources for 3D connections. This

impact worsens at advanced technology nodes due to higher congestion of interconnects. The increase in signal wirelength translates into additional net switching power dissipation, which significantly contributes to total power. This in turn aggravates thermal issues further in 3D ICs. In this work, a comprehensive study on the impact of power delivery network (PDN) on full-chip wirelength, routability, power, and thermal effects in gate-level monolithic 3D ICs across different technology nodes is presented. The thermal impact of PDN is also examined. Next, various PDN design optimization techniques for monolithic 3D ICs are proposed to minimize switching power overhead under the given IR drop budget.

- **CAD solutions to handle low device performance in top-tier of monolithic 3D IC:** For fair assessment of monolithic 3D ICs, the designs should be optimized and evaluated under practical settings, which include the impact of low performance transistors in top-tier due to low thermal budget during fabrication. Also the design techniques used need to be similar or better than state-of-the art commercial RTL-GDSII flows, while handling these issues. This work studies the system level design impact of low performance devices in top-tier of M3D with existing flows and then develops a new CAD solution to handle such device issues during physical design, while minimizing design and runtime overhead.

- **Tier partitioning strategies to mitigate back-end-of-line (BEOL) impact in monolithic 3D ICs** Tungsten has to be used in the bottom-tier of M3D since copper cannot handle the temperature used for fabrication of the sequential device layers. Tungsten is more than three times resistive than copper and affects the timing closure and performance quality of the M3D designs. In addition, 3D routing is not negligible, especially in advanced nodes. This research dissertation studied and quantifies the adverse impact of BEOL in in the bottom-tier of monolithic 3D ICs. An effective path-based partitioning methodology is developed to mitigate the performance

degradation due to tungsten without any additional design optimization. Next, a net-based tier-partitioning methodology is developed to reduce congestion and metal layer requirement in bottom-tier of monolithic 3D IC and hence save on BEOL cost.

- **Monolithic 3D IC power, performance and cost comparison:** As monolithic 3D IC is being carefully studied and evaluated as a feasible alternative to scaling or an extension to an existing technology node, a detailed power, performance, and cost analysis is imperative. Any future generation technology node requires reduction in power, savings in cost, and improvement in performance. In this work, a comprehensive study of power, performance, area, and cost comparisons among TSVs, mini-TSVs (TSV with smaller diameter), and MIVs is presented. In addition, the magnitude of power savings of M3D is heavily dependent on the selection of device technology and process design kit (PDK). Therefore, the impact of transistor technology on the power savings in monolithic 3D ICs is also discussed.

1.3 Organization

The rest of this dissertation is organized as follows:

- In Chapter 2, thermal modeling and optimization of monolithic 3D ICs is presented.
- In Chapter 3, full chip power delivery impact on monolithic 3D ICs and related optimization techniques are presented.
- In Chapter 4, CAD solutions for low performance transistors in the top-tier is presented in the first section. Tier partitioning strategies to mitigate tungsten and BEOL impact in bottom tier is presented in the second section
- In Chapter 5, power, performance and cost analysis of monolithic 3D ICs is present along with comparison across device technologies and with TSV-based 3D ICs and 2D ICs.

- In Chapter 6, additional topic of near-threshold voltage 3D IC design is presented.
- In Chapter 7, the conclusions of this dissertation are summarized, along with the discussion on the possible future works.

CHAPTER II

THERMAL MODELING AND OPTIMIZATION OF FULL-CHIP MONOLITHIC 3D ICS

Thermal issues is one of the biggest challenges with 3D ICs in general. Monolithic 3D integration is a fairly new technology and its thermal properties have not been studied well. In addition, the high integration density in M3D aggravates the power density and temperature issues further. This makes thermal-aware design of M3D a very important and practical neccessity.

In this chapter, the thermal properties of M3D technology is studied and a fast-accurate temperature model is developed. This model is then used inside a 3D floorplanner to carry out thermal-aware floorplanning without incurring any design runtime overhead. The modeling and floorplanning is carried out for different package structures.

2.1 Motivation and Background

Monolithic 3D ICs enable power savings over 2D ICs by reduction of interconnects and associated cell-usage. However, the increased power density affects the temperature of the chip, especially in the layers away from the heat sink or other equivalent cooling features in modern miniaturized electronics [51, 22]. Therefore, importance of thermal-aware design methodologies become more critical in 3D ICs. The major bottleneck of considering thermal aspect within the physical design process is the huge runtime required for accurate temperature analysis. The inclusion of such detailed analysis within the design process is not practically feasible.

There exists several works which focus on the thermal issues and thermal aware design of TSV based 3D ICs [16, 17]. Prior works have tried to develop accurate temperature

evaluation models to be included within the chip design process [26]. The use of compact resistive thermal grid network to estimate the temperature profile of a chip has been studied by Cong *et al.* [17]. Compact resistive model and hybrid model within the floorplanning process is used to analyze the temperature and insert whitespace for dummy vias. The calculation of resistive network solving still consumes some runtime and the insertion of whitespace increases the area further, diminishing the 3D IC benefits. 56% reduction in temperature is reported but with a large area increase of 21%. The optimization of silicon area is important in 3D ICs along with the temperature rise and too much area cannot be sacrificed for temperature improvement. Zhou *et al.* [58] propose a force-directed floorplanner approach to spread high power blocks while simultaneously optimizing wirelength, area and thermal distribution. The modeling of temperature based on total leakage power dissipation and its use in the tier-planning of similar layout processor chips is demonstrated by Juan *et al.* [29]. The 3D overlap estimation along with power density calculations for thermal-aware planning has been used in [27].

In case of compact thermal modeling, there have been many studies on for both physical design as well as for model predictive controllers (MPC) to have real-time thermal management in place of conservative worse case thermal management for multi-core chips. Compact thermal modeling for realistic energy-aware thermal management and control techniques with proper validation for multicore chips has been done in [7]. A robust thermal model using graybox approach is developed in this work, It uses both statistical content and physical laws for better quality. Their adaptive models are used as controllers during operation and cover 2D multi core designs. Hotspot tool [26] is one of the most popular thermal analysis tools widely used in research. Compact resistive models with different tuning parameters for trade-off between run-time and accuracy are used. The grid model is capable of handling 3D stacked chips with different power dissipating layers in the compact resistive mesh. Beneventi *et al.* [8] developed a compact thermal model for TSV-based logic+WideIO 3D stack. This model can successfully predict the temperature at locations

where sensors are absent and can also evaluate the power dissipation based on temperature data.

All the above works on 3D IC thermal modeling cover TSV-based 3D ICs only and involve various forms of matrix manipulation, which though simplified, is still computationally expensive. The incur extra runtime or use indirect methods of thermal analysis. In addition, only conventional package and stack up (with a heat sink at the top) is covered in these works.

Interestingly, monolithic 3D ICs exhibit different thermal behavior due to their layer structure and are not as thermally bad as TSV-based 3D ICs even though copper TSVs increase conductivity. These unique properties facilitates the development of a very fast temperature model with high degree of accuracy. In addition to that, 3D ICs also provide huge potential in the design of low power processors for use in mobile applications and monolithic 3D ICs specifically enable ultra high packing density [36]. However, mobile applications have different package structure due to their size and weight constraints. Heat sink is absent in such packages and different materials are used for spreading and dissipation of heat. Therefore, to tap all benefits of monolithic 3D IC to the full extent, it is very important to also take into account the different types of package structures during thermal optimization, because they will significantly impact the overall thermal quality. The thermal model and floorplanner developed in this work is the first to cover monolithic 3D ICs and is a fast and simple using very few input parameters.

2.2 New Issues and Unique Thermal Properties of Monolithic 3D ICs

2.2.1 Monolithic 3D integration

Monolithic 3D integration technology enables ultra high density vertical integration. The advanced manufacturing technology allows active device layers as thin as 10nm to be integrated over one another with high alignment precision [5]. To understand the thermal properties of monolithic 3D ICs, a good understanding of the structural details is essential.

A typical two-tier monolithic stackup is shown in Figure 2a in a flip-chip configuration. The first set of transistors closer to the handle bulk are processed with standard SOI process and make up Tier 1. A thin inter-layer dielectric (ILD) is deposited over the metal layers for the bonding of the next device layer. This device layer along with the metal layers make up the other tier (Tier 0) of the 3D stackup. The transistors in these layers are processed with low temperature process ($<$650 ^oC). Since this chapter focuses on thermal modeling and floorplanning, it is assumed that devices in all tiers of monolithic 3D IC have similar performance [3]. The performance difference issue is addressed in a later chapter. Also for this chapter, the tier numbering convention in 3D ICs is such that Tier0 (bottom tier) is the one closest to the printed circuit board (PCB) and the tier numbers increase on going further away from the PCB.

2.2.2 Material and structural differences

The differences in fabrication process of monolithic 3D and TSV-based 3D result in significant differences in their thermal behavior. Figure 2 highlights differences in the materials used in the two technologies. Their conductivity and thickness influences the thermal behavior. Table 1 lists their details for a typical 45nm technology process. The relative contribution of each material per tier is also shown in the table.

In TSV-based 3D ICs, copper TSVs and μ-bumps improve the conductivity. However, the presence of bonding layer (underfill) which is necessary for stress related issues worsens the overall conductivity significantly (Figure 2b). Typical materials used for underfill are required to be soft and elastic and in general such materials have poor thermal conductivity. BCB is one of the commonly used materials and it has a thermal conductivity of 0.29 W/m-K. Copper metal on the other hand has a thermal conductivity of 401 W/m-K. The presence of this underfill which is around 2.5μm thick impedes the heat flow from Tier0 towards the heat sink present above the handle bulk resulting in considerable temperature rise in Tier0. However, heat from Tier0 passes through silicon substrate before reaching the underfill

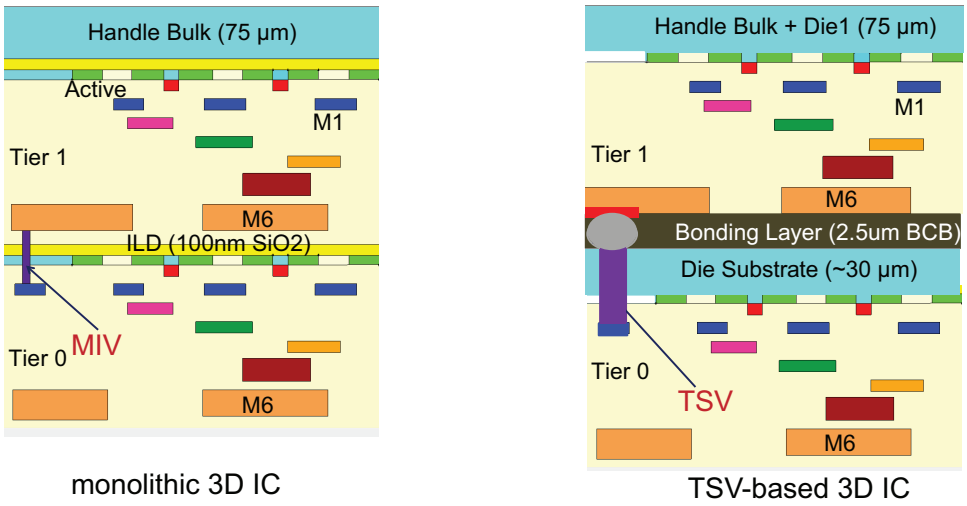


Figure 2: 2-tier 3D IC layer structure (heat sink on top) of monolithic 3D IC vs TSV-based 3D IC

wall. Silicon being a good conductor of heat spreads out the thermal profile of Tier0 by allowing many lateral heat flow paths in its $30\mu m$ thickness. Tier1 in TSV-based 3D ICs does not have any buried oxide between the device layer and the handle bulk. This helps in better conduction of heat from Tier1 to the heat sink.

In contrast to TSV-based 3D ICs, the bonding layer and bulk substrate are absent in monolithic 3D ICs, while the different tiers are separated by inter-layer dielectrics (ILD) which function as the buried oxide for the SOI process for formation of subsequent device layers. Also the MIVs are tiny compared to the huge TSVs. These particular differences change the heat dissipation phenomenon of monolithic 3D ICs from that of TSV-based 3D ICs. The absence of bulk substrate and the extremely thin device layers reduce the lateral conductivity to almost zero which results in heavy tier-to-tier thermal coupling. The heat flows only vertically up until it reaches the handle bulk where there is lateral spreading due to its very large thickness compared to all other layers. The presence of buried oxide also increases the thermal resistance from top tier to handle bulk. All these factors considered together result in similar temperature profiles for all the tiers irrespective of the whitespace locations in the different tiers. A high power block in the tier closer to the heat sink will also result in a hot spot in all other tiers away from the heat sink. There is a difference in

Table 1: The different materials used in the layers, their thermal conductivities, vertical thicknesses and relative % in total stack

Layer/Structure	Material	Thermal Cond. (W/m-K)	Vertical Thickness	% of total Tier0	% of total Tier1
Monolithic					
Handle Bulk	Silicon	141	$75\mu m$	-	97.1
ILD (Inter-tier)	SiO_2	1.38	100nm	4.3	0.13
BEOL	SiO_2/Cu	1.38/401	$2.2\mu m$	93.6	2.84
TSV-based					
Handle Bulk+Die 1	Silicon	141	$75\mu m$	-	97.2
Die0 Substrate	Silicon	141	$30\mu m$	86.5	-
Bonding Layer	BCB	0.29	2.5 μm	7.2	-
TSV	Copper	401	$30\mu m$	in Die0 sub	
TSV-bump	Solder	50	2.5 μm	in BCB	
BEOL	SiO_2/Cu	1.38/401	$2.2\mu m$	6.3	2.9

the temperature value of the same 2D location in two tiers due the rise across the 100nm ILD. Also the maximum temperature of the tier closest to the heat sink is more than that of TSV-based 3D IC due to the presence of additional oxide layer which is a poor conductor.

2.2.3 Vertical tier to tier coupling in monolithic 3D ICs

Figure 3a shows the layouts of a 3-tier monolithic block level 256 bit multiplier. Figure 3b are the temperature maps of the individual tiers with 2D thermal analysis performed on each tier independently for conventional package with heat sink. The cooler regions (blue) and their spread exactly following the whitespace locations in the corresponding layouts of Figure 3a. Since it is 2D thermal analysis, these whitespace locations have no heat generation and hence the cooler spots. However, with 3D thermal analysis for this 3-tier 3D design considered as a whole, the temperature maps change significantly (Figure 3c). The hotspots of all the individual tiers overlap with each other in 3D and affect the temperature of all the three tiers with bottom tier suffering the most due to additional poor conducting ILDs on the way to the heat sink. The hotspots (red color) in Tier2 affect the 3D design most because it obstructs direct vertical heat flow from the tiers lying below along with addition of its own heat. In M3D, there is negligible lateral conduction until the heat reaches the

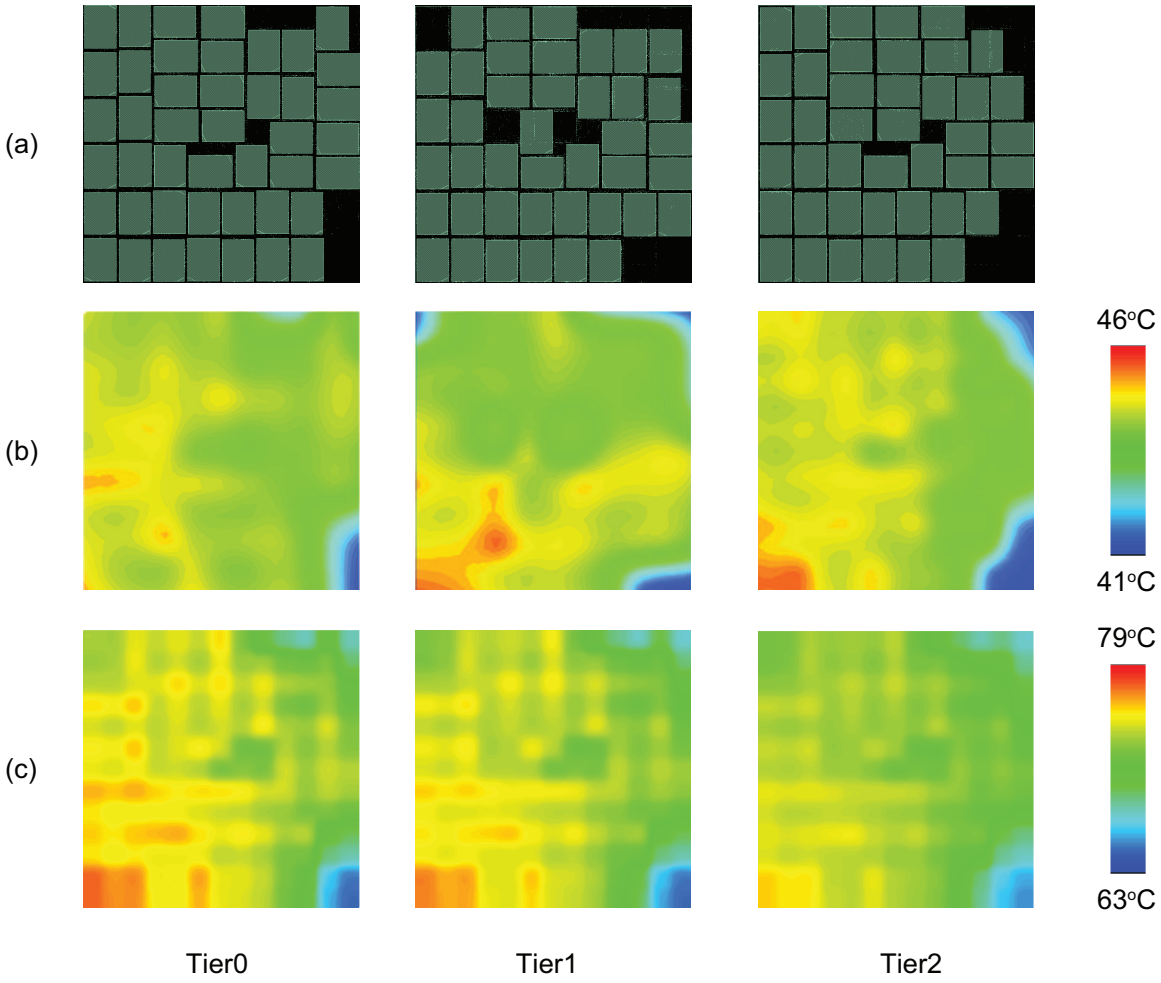


Figure 3: Tier to tier coupling in monolithic 3D ICs. (a) 3D-floorplan for 3-tier 256-bit multiplier (b) Temperature maps with independent 2D thermal analysis of each tier (c) Temperature maps with stacked 3D thermal analysis

handle bulk. Therefore, the temperature maps are similar in trend of variation across the entire area. The temperature values also increase compared to the individual 2D analysis of each tier due to almost three-fold increase in power density. Only the common whitespace regions (bottom right corner) remain cooler in all the tiers.

2.2.4 Temperature map comparisons

Figure 4 shows the temperature map of a same 2-tier 3D layout in monolithic technology and TSV-based technology. These temperature maps for the two technologies are compared along the lines of the discussions presented earlier and highlight the unique properties in

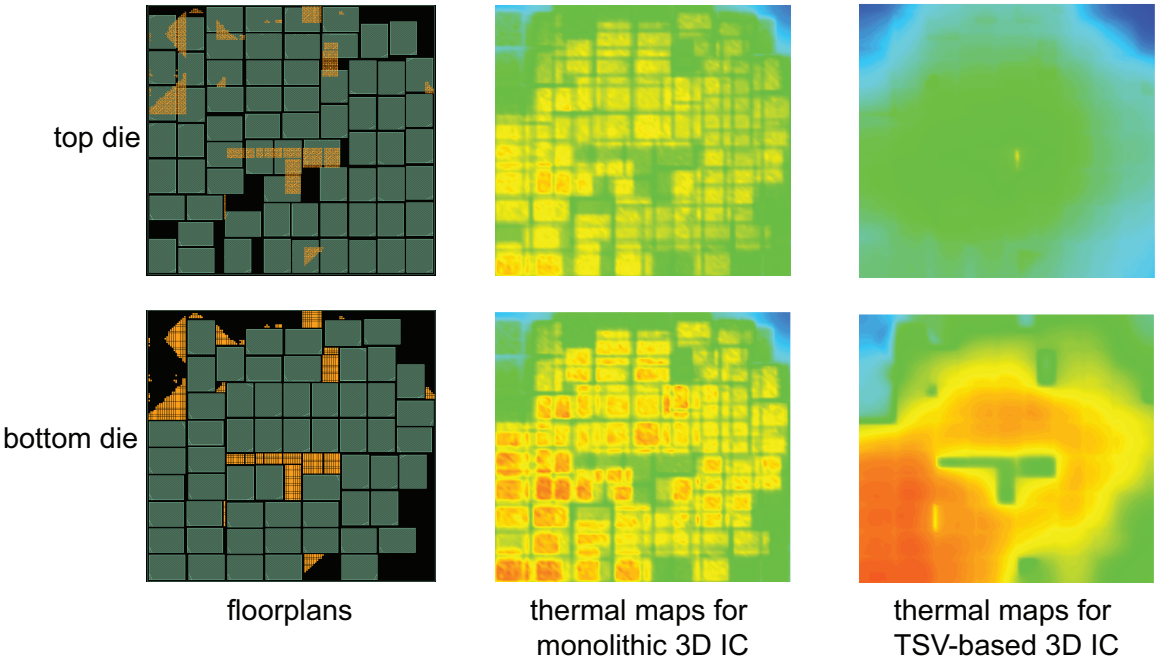


Figure 4: Temperature maps of same 2-tier 3D floorplan (originally designed for TSV-based 3D IC) in monolithic 3D IC technology and TSV-based 3D technology. The temperature range is [61°C, 71°C].

monolithic 3D ICs.

The layout is originally designed for a 2-tier TSV-based 256 bit multiplier. The TSV locations are shown in yellow in Tier0 layout and their landing pads shown in Tier1. Since the primary objective here is to understand the thermal behavior of the technology, for fair comparison from the thermal point of view, the same 3D layout with same power density and performance is analyzed for a monolithic structure with TSVs replaced by MIVs at the same 3D via locations. In practice, MIVs are much smaller and their design will consume much lesser area.

For the TSV-based 3D IC temperature map, the presence of TSVs help in improving the conduction significantly in Tier0. There are cooler spots among very hot ones wherever TSVs are present. The temperature of other regions is quite high due to the heat flow obstruction by the bonding layer. Tier1 is much cooler compared to Tier0 as it is closer to the heat sink. The other important thing to observe is the lateral spreading of temperature across the two tiers which smears the temperature profile of each tier. This is because of

the bulk silicon substrate which allows multiple lateral heat flow paths.

For monolithic 3D IC design on the other hand, the temperature profiles of the two tiers are identical and the block layouts can be demarcated in the temperature map itself. This is a result of absence of lateral conduction at the source of power dissipation. The vertical tier-to-tier coupling can be observed by the block outlines from both tiers appearing overlapped in the temperature maps. Tier0 map is hotter than Tier1 due to the heat block by the ILD. Tier1 of TSV-based 3D IC is cooler than Tier1 of monolithic because of the absence of oxide which is a poor thermal conductor. Tier0 in TSV-based 3D is much hotter than Tier0 in monolithic 3D due to bonding layer which is a poorer conductor than SiO_2. Wei *et al.* also compared TSV based 3D IC with monolithic 3D ICs but did not consider the underfill layer [54]. The mass production of TSV-based 3D ICs without any underfill is highly unlikely due to stress related issues. Therefore, it is important to consider them during thermal behavior study of TSV-based 3D ICs and then compare with monolithic 3D ICs. This very poor conducting bonding layer in TSV-based 3D ICs significantly worsen the temperature of tiers away from heat sink. If this layer is ignored during analysis, then TSV-based 3D ICs will be better than monolithic 3D ICs thermally.

The key points from the above thermal study of monolithic 3D ICs and comparison with TSV-based 3D ICs are

- Monolithic 3D ICs have almost zero lateral conduction at the source of power due to very thin layers and show no lateral spreading in the device layers.
- There is heavy vertical tier-to-tier coupling in monolithic 3D and all tiers have similar temperature profile with differing absolute values due to rise across ILDs.
- In monolithic 3D ICs, handle bulk is the first layer in the path of heat flow where noticeable lateral conduction occurs. Therefore, the individual neighbors in a floor-plan have an indirect effect unlike TSV-based 3D ICs where they directly affect each other by conduction through silicon substrate.

- MIVs do not play an important role in heat conduction like TSVs due to small size and thickness.

2.3 New Mobile Package Structure and Properties

Miniaturization is one of the key characteristics of modern VLSI. With low power devices like smart phones, smart watches, sensor nodes etc., there is a need for compactness and light weight materials and high integration density. The power dissipation in such applications is much lower than that of high-performance servers and desktop computers. Large and heavy heat sinks with cooling fans can be avoided for such systems. Therefore, industry uses a different kind of packaging structure for the integrated circuits used in mobile applications. Figure 5b shows the structure and materials used for packaging and cooling of mobile processors [44].

Since monolithic 3D ICs enable very high integration density, they are a very good candidate for use in mobile processors to increase functions in the same form factor. Such mobile systems use the new mobile package structure and there is a need for good thermal planning and budgeting for use of monolithic 3D ICs. This is the key motivation to study mobile package structure in detail, analyze the properties and impact of the new materials used and develop thermal model which can incorporate the package characteristics during fast accurate temperature evaluation. Furthermore, knowing the impact of various materials used in the package will enable designers and packaging engineers to carry out package optimization after thermal optimization during physical design. This section give an overview of the the package structure used in mobile phones. The differences in the cooling phenomenon for such packages in contrast to conventional packages with heat sink and dependence on number of tiers in M3D is also presented.

2.3.1 Structure and materials

Figure 5 shows the package structure for conventional cooling with heat sink (Figure 5a) and mobile applications without any heat sink (Figure 5b). The major differences in the

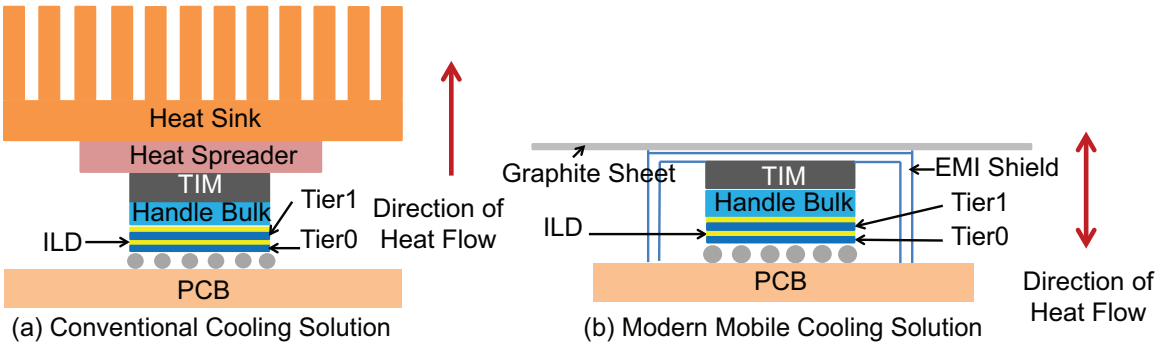


Figure 5: 3D IC packaging structure for cooling (a) conventional cooling (with heat sink) and (b) modern mobile cooling (no heat sink)

Table 2: Properties of the different layers in mobile package structure.

Layer	Vertical thickness (μm)	Therm. cond. (W/mK)	
		Vertical	Lateral
PCB	800-1500	1.5-4.5	25-60
Handle Bulk	50-200	141	141
Therm. Int. Material (TIM)	500-1200	0.5-5	0.5-5
EMI Shield (Steel/Al)	100-250	20/120	20-120
Graphite Sheet	25	2.9-4.5	300-500

mobile package is the absence of a copper heat sink with multiple fins and copper heat spreader.

The absence of heat sink with multiple fins and cooling fan in mobile package structure reduces the dominance of the upward path in heat conduction. The red arrows in Figure 5 show the primary direction of heat flow in the respective structures. Because of a very large heat sink and low thermal resistive path, almost all of the heat flows towards the heat sink in conventional packages. However, for mobile packages, heat flows in both directions and therefore, the importance of all other layers increase. The different layers in the mobile package structure and their thickness and conductivity values are shown in Table 2. The values are shown in ranges as the properties of some of the layers can be different in different systems, based on the actual composition and requirement.

Along with the printed circuit board (PCB) inside the mobile phone towards the display side, the back body also helps in heat dissipation and the very thin graphite sheet helps in

spreading the heat to the entire back cover instead of having concentrated hot spots. The electromagnetic insulator (EMI) which is usually a steel or aluminum sheet also helps in spreading of the heat by providing a low heat resistive path along with its primary function of shielding. Another important factor is that the lateral conductivity of PCB and graphite sheet are much better than their vertical conductivities. Therefore, they play a significant role in good lateral spreading and hence increasing the surface area of contact with the external environment. The same design with same power maps will have hotter spots with mobile package than a package with heat sink. However, mobile processors have significantly reduced average power density (< 2 W/cm^2) compared to high performance servers (20-30 W/cm^2) and hence the maximum temperature is well within control even with the absence of heat sink and fan. In this work on thermal analysis related to mobile packages, the PCB, the graphite sheet and the free regions of EMI are all connected to ambient environment.

2.3.2 Comparison with conventional package structure

The absence of heat sink significantly changes the thermal behavior of mobile packages in contrast to that of conventional packages with heat sink. This difference becomes more prominent in multi-tier 3D ICs where there are multiple layers of heat source. Figure 6a shows the block level layout of 3-tier OpenSPARC T2 core [43] with the highest power density execution unit blocks highlighted in red outline. The floorplan is targeted towards minimum wirelength. Figure 6b and Figure 6c are the temperature maps with the conventional and mobile package structures respectively. The total power dissipation values of the system under the two packaging structures are different, with the mobile package having much lower power density. The temperature ranges are normalized with blue color for minimum and deep red for maximum temperature in the respective packages.

As discussed above, almost all of the heat flows towards the heat sink in conventional packages. This makes the tier away from the heat sink most critical in terms of thermal

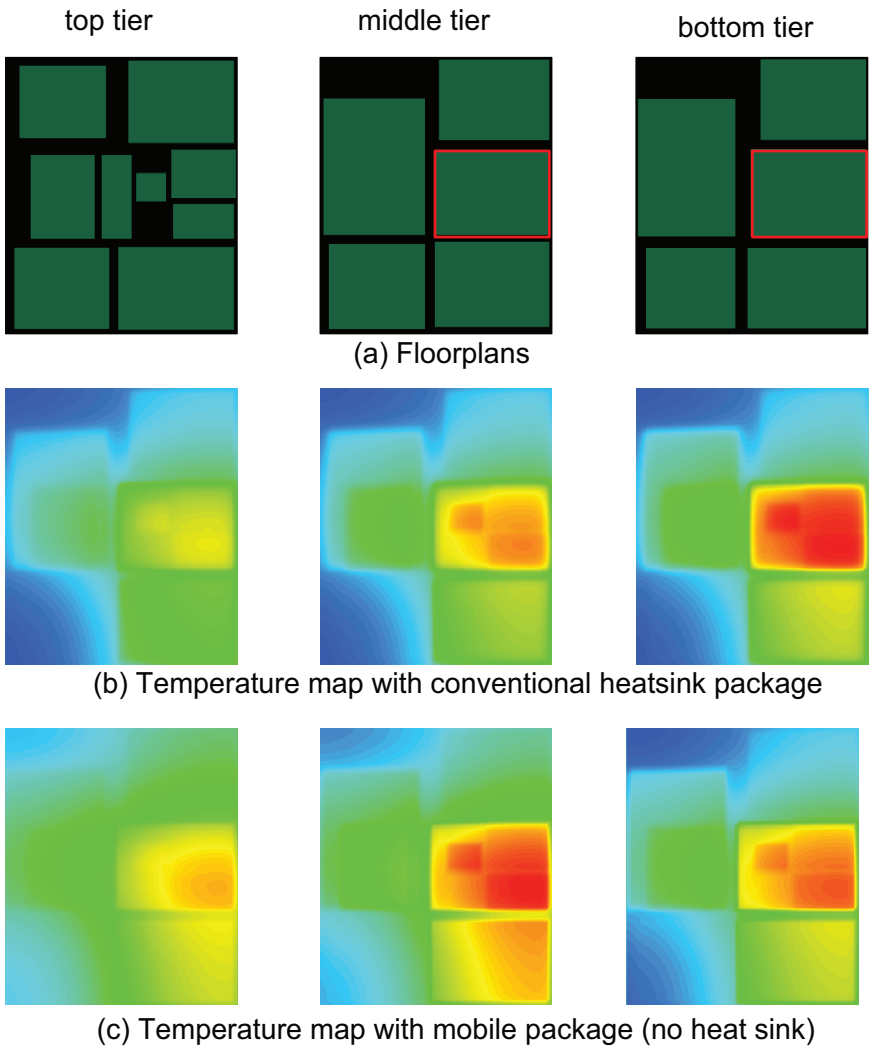


Figure 6: Temperature hotspot and distribution comparison for conventional cooling (with heat sink) and modern mobile cooling (no heat sink) for openSPARC T2 core. Temperature scales are normalized with blue as minimum and red as maximum temperature for respective package structure with lower power density for mobile systems. Red outline blocks in (a) are the maximum power density blocks.

reliability. This is evident from the red hot spots in the bottom tier in Figure 6b which is farthest from heat sink. On the other hand, bi-directional heat-flow in mobile packages has two important consequences. First is that the middle tier is most critical in terms of thermal reliability unlike conventional packages (Figure 6c). Secondly, due to bi-directional flow, the relative temperature difference between tiers is lesser compared to conventional package structure. This difference can be seen in the temperature maps of Figure 6 where the

Table 3: Maximum temperature rise values (above room) across different tiers in designs with different packages.

Design	Conventional Package			Mobile Package		
	tier0	tier1	tier2	tier0	tier1	tier2
2D IC	24.37	-	-	16.54	-	-
2-tier 3D IC	47.10	44.30	-	32.27	31.60	-
3-tier 3D IC	65.82	63.10	59.95	43.18	44.15	42.4

relative difference in maximum temperature across tiers is lesser for mobile packages and more for conventional packages.

2.3.3 Thermal behavior with different number of tiers

Since heat flow is bi-directional in mobile packages, the middle tiers are more critical for multi-tier 3D ICs and the extreme tiers are influenced similarly. This implies that 2-tier 3D ICs are almost similar to 2D ICs in terms of thermal floorplanning for same mobile package properties, unless the power map is heavily unbalanced to have excessive power dissipation on one tier only. A high power density block can be placed in either of the two tiers in a 2-tier 3D IC design to have the same overall temperature profile because heat flows in both directions almost equally. This is not the case for conventional packages with heat sink because the tier away from the heat sink is always more critical thermally and it is desirable to have the high power density blocks closer to the heat sink. Table 3 shows the maximum temperature of different tiers for a 2D, 2-tier 3D and 3-tier 3D design with same total power. The maximum temperature for the two tiers in 2-tier 3D IC is almost same for both tiers with mobile package but for 3-tier case, the middle tier has higher maximum temperature than both the extreme tiers. The increase is uniform for conventional package with the tier away from the heat sink having worst temperature in all 3D designs.

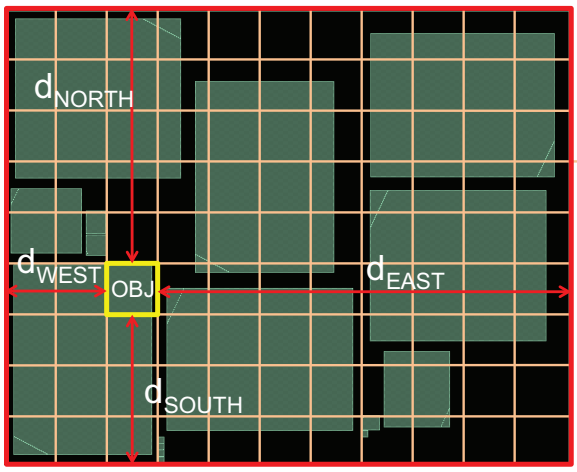


Figure 7: Final model structure with an objective tile. The red rectangles show the objective tile and rest of the chip. Their power values along with 2D distances from boundary are used as inputs for temperature calculation

2.4 Fast Thermal Analysis Model

2.4.1 Model development

Steady-state finite element thermal analysis leads to large matrix calculations of an equivalent thermal resistive network with multiple power sources. Therefore, a non-linear regression based technique is used to accurately model the steady-state temperature of monolithic 3D ICs after generating a large number of representative samples. The method of approximating a quantity dependent on certain number of predictor inputs using such techniques has been used in prior works [32]. While regression helps in determining direct correlation between target and inputs, non-linearity helps is reducing the total number of required inputs without affecting prediction accuracy. Temperature is set as the target quantity and it is modeled after successfully determining the different parameters of the monolithic 3D IC on which it depends. The developed temperature model evaluates the steady-state thermal behavior of monolithic 3D ICs of given dimensions, number of tiers and power distribution.

2.4.1.1 Initial experiments

The entire chip is divided into a tile based structure for each tier as shown in Figure 7. Each of the tiles is randomly assigned a power value such that the power density lies between

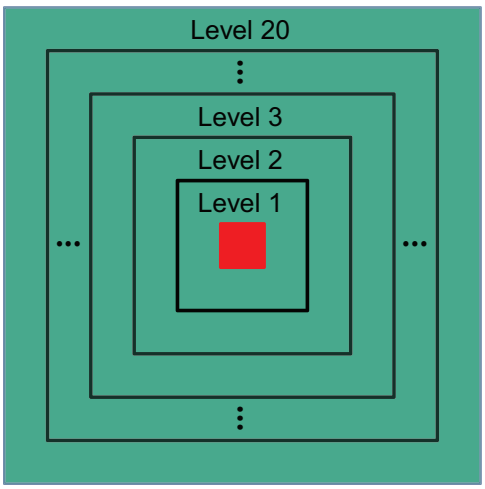


Figure 8: Experimental setup with 20 neighboring levels and objective at the center.

0-100W/cm^2. Full chip thermal FEA with 20μm X 20μm mesh is carried out on these test cases. To address different types of applications,two kinds of packaging structure are considered independently. Figure 5a is the conventional cooling method which uses heat spreader and heat sink. Almost 100% of the heat dissipates through the heat sink. Figure 5b is the packaging structure used in modern smart phones due to size limitations [44]. The thermal resistance in both directions is similar therefore, there is bidirectional heat dissipation. ANSYS Fluent and Spice simulations are used to carry out the thermal analysis [2].

Based on the study of the thermal properties presented earlier,various experiments are conducted involving different power distributions, different granularity of tile division, multiple neighboring levels considered separately vs considered as single entity lumped together and temperature dependence on 2D and 3D location of the objective tile. The neighboring tiles of an objective tile were found to have a unified effect on the temperature of the objective. The reason is that they affect the objective indirectly through the handle bulk and not directly because immediate lateral conduction is almost absent. It is also observed that the location of a particular tile in the layout affects its temperature value. Some of the experiments conducted are explained below.

Table 4: Experimental results with different number of neighbors considered during MARS modeling

No. of levels considered	GCV	Avg Error (%)	RMSE	Most important variable
20	0.108	1.31	0.46	Power_ Level20
19	3.855	8.04	2.66	Power_ Level19
18	6.550	11.26	3.90	Power_ Level18
17	7.475	13.66	4.73	Power_ Level17

Table 5: Experimental results of modeling with the entire chip area considered completely but with different number of partitioning levels

No. of partitions	GCV	Avg Error	Most important variable
20	0.108	1.31	Power_ Level20
10	0.105	1.53	Power_ Level10
5	0.199	1.77	Power_ Level5
4	0.20	1.95	Power_ Level4
2	0.626	2.14	Power_ Level2
1	0.727	2.32	Power_ Level1

The primary goal is to divide the entire chip into tiles of $100\mu m$ X $100\mu m$ and then obtain a model with minimum number of inputs to calculate the temperature of each tile accurately without carrying out full FEA simulations. The lesser the number of input variables required, faster is the full-chip temperature analysis. To correctly determine the number neighboring levels to be covered, experiments were done starting with 20 levels of neighboring tile levels and dropping the farthest neighbor one at a time to study the effects on the results (Figure 8). Since each tile is $100\mu m$ X $100\mu m$ and there are 20 levels of neighbor rings, the chip size is 4.1mm x 4.1mm. The effectiveness of a model is measured in terms of the generalized cross validation (GCV) values of the model development and average error. A GCV value close to zero implies perfect modeling. The results of this experiment show that the amount of error increases as the farthest neighbors are removed (Table 4). Therefore, the farthest neighbors have a considerable effect on the temperature of the objective even though they are far away laterally. This is because the total power of the larger rings of tiles are much more than the objective and all of this heat goes vertically to the handle bulk layer before indirectly affecting the objective tile temperature. Therefore, any

power dissipation cannot be ignored, irrespective of the lateral distance from the objective tile.

Using the same raw data as in the sub-section above, an analysis is carried out from a different viewpoint. The entire region (20 levels) is divided into different number of equal regions viz. 20 partitions (default), 10 partitions, 5 partitions, 4 partitions, 2 partitions and finally a single partitions where all 20 neighboring tile rings are treated as one. The power dissipation of these different partitions are used as variables to develop the model and the resulting model is compared in terms of GCV and average absolute error (Table 5). The results show that it is not necessary to have fine grained neighbors in the model. All the neighbors near or far have similar effect. Once again, this is explained by the indirect effect of neighbors through the handle bulk which is $75\mu m$ thick and is silicon. The most important variable is always the last partition which has maximum magnitude of power.

2.4.1.2 Modeling technique

From the above experiments, the following important parameters which influence the chip temperature are finalized.

- Power of objective tile
- Total Power of rest of the tiles in the same tier
- Lumped sum of power of all tiles exactly above the objective
- Lumped sum of power of rest of tiles of the above tiers (excluding the ones directly above)
- Lumped sum of power of all tiles exactly below the objective
- Lumped sum of power of rest of tiles of the tiers below (excluding the ones directly below)
- Distance of the tile from each of the four 2D boundaries (4 variables)

- Distance from vertical boundaries (3D location).

The contributions of all power values other than that of the objective and immediate vertical neighbors can be summed up because all lateral influence is indirect through lateral conduction at the handle bulk only, which is above all the device layers (Figure 2a). The exponential increase in leakage with temperature can be taken care of by separating the power inputs into its components viz. dynamic and leakage powers and updating the leakage powers with temperature increase till a specified tolerance level is met.

Figure 7 shows the division of chip and the 2D dimension related variables. The target variable of the model is the rise in temperature above room temperature. Modeling is carried out with the help of Multivariate Adaptive Regression Splines (MARS) which is a non-linear regression technique [50]. The number of inputs is minimized to keep the final temperature evaluation runtime less but with very good accuracy. The chip dimensions are implicitly taken care of by the distance variables and are excluded in the inputs. The tier number of the objective is also included to include the 3D distance from the package boundaries. The individual tile size is fixed at $100\mu m$ X $100\mu m$. Further granularity does not improve modeling results much but adds to the evaluation time for the whole chip which will affect the overall runtime of the thermal-aware floorplanner presented later. The thermal analysis models are developed for each of the packaging structures separately for both 2-tier and 3-tier 3D cases.

2.4.1.3 Sample generation

A large number of samples which cover all the possible variations in the parameters are required to develop a good model. To correctly capture all the possible 3D chip sizes and power distributions, detailed thermal analysis of whole chip testcases is carried out. These cases cover chip dimensions from 1mm to 5mm (in steps of 1mm) with aspect ratio lying between 0.5 and 2. Each chip is divided into $100\mu m$ X $100\mu m$ tiles and each such tile forms one sample. The above properties add up to 17 whole chip FEA simulations. These

simulations are run only for one time to generate a large number of samples. Power density values of the tiles are randomly distributed from 0-100 W/cm^2 while ensuring that around 10% of the total chip area is whitespace to correctly simulate practical designs. Around 15% of the samples are used for training of model and the rest used for testing. Since the samples were generated with the respective packages, the training captures the package properties into the final temperature model. For a different package structure or same package with different dimensions or material properties, the training samples generated with the corresponding package will capture the package properties. Therefore, the same modeling approach adapts the model to the package used for generating the training samples. During whole chip thermal simulation to obtain these samples, Back End of Line (BEOL) material is treated as 100% dielectric (SiO_2) material. This is because these generated samples do not have actual routing and dielectric constitutes maximum portion of BEOL [2].

The modeling results are more accurate when all the samples have a random power density distribution with fixed average rather than with varying average. Therefore, samples with power density varying randomly from 0-100W/cm^2 are used, which results in an average power density of all samples close to 50W/cm^2 (average of a random distribution). However, the power density of the practical case to be modeled will vary from design to design and needs to be taken care of during final evaluation. The trend prediction of the developed model is always correct, irrespective of the actual average power density. However, the values are just shifted up or down and need a constant correction offset depending on the actual power density being greater than or less than 50W/cm^2. From various practical example cases, this offset is evaluated to be a constant multiple of the difference of the actual average power density (PD) of chip and the samples' power density (=50 W/cm^2 here). To successfully model samples covering different average power density, the number of total samples required increase by orders of magnitude. Since steady state average temperature is a linear function of average power, this simple offset method avoids the need

for generating more samples for modeling. The exact multiplying coefficient depends on the samples used for modeling but will always remain constant once a model is developed irrespective of the actual chip being evaluated. The temperature evaluated by the model is the rise above room temperature as it is the more appropriate variable to model. To get the absolute temperature, the room temperature can be added to it.

2.4.2 Model accuracy

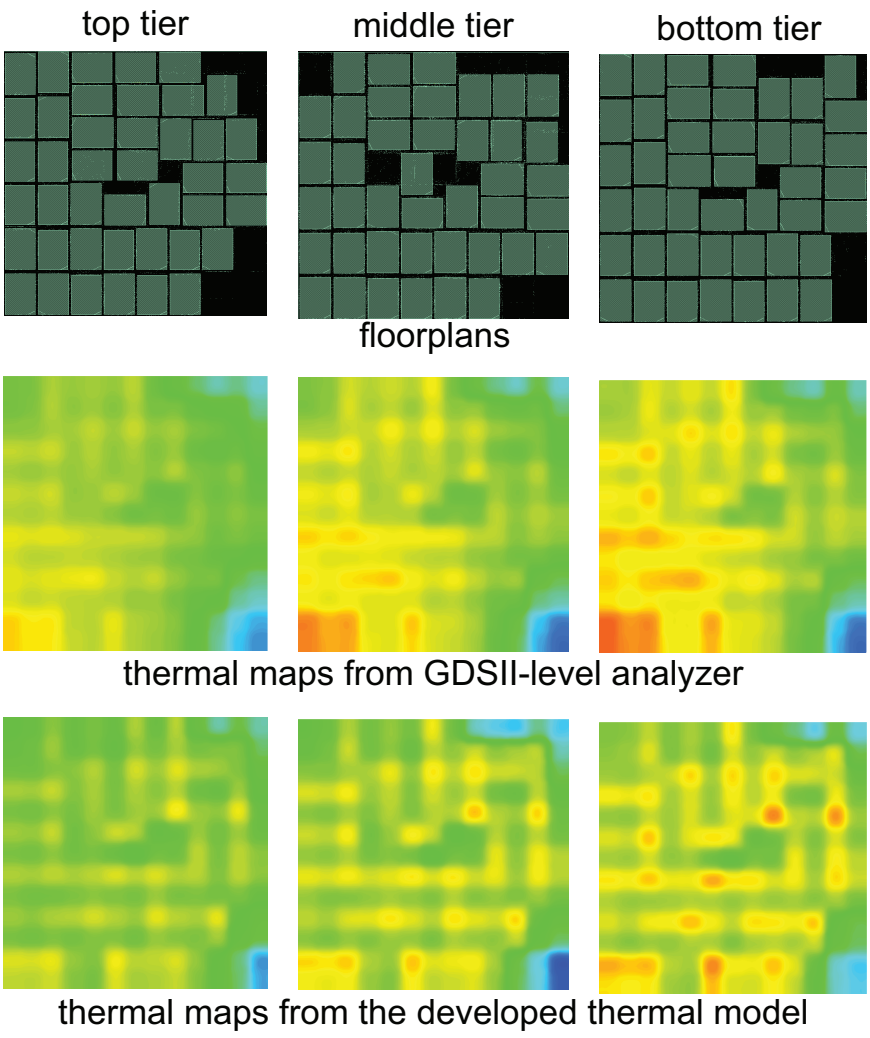


Figure 9: Model accuracy: FEA simulation vs new temperature model for 256-bit multiplier. The temperature range is [63°C, 79°C].

The testing sample set gives an absolute average error of less than 1%. For practical designs, the average error is less than 5%. Figure 9 shows the accuracy of model for a

Table 6: Full chip thermal analysis runtime comparison for 3-tier 3D IC (1.3 mm x 1.3 mm footprint). (Runtime for new model averaged over 10^6 runs)

Method	Runtime (in sec)	Normalized Runtime
New Model	0.00022	**1.0**
GDS-level FEA	1082	**4.9 x 10^6**
HotSpot	5.68	**2.6 x 10^4**

testcase, designed for 3-tier 3D. The top row show the layouts of the individual tiers of 3-tier 3D IC, the middle row is the temperature maps after detailed FEA thermal analysis with the average temperature of each tile plotted while the last row is the temperature analysis results from the developed model. The model captures the temperature variation trend very well and all the hotspots are accurately detected. This methodology of temperature estimation can be used for any circuit irrespective of whether it is a flat gate level design or a block level design. Only the power needs to be distributed into the tiles to carry out fast accurate temperature analysis.

The important conclusion is that irrespective of the error, the trend of temperature change within the chip is accurately evaluated with the developed model. The error primarily comes in the cooler regions of the chip. The reason is that there is immediate whitespace on all sides (2D and 3D) around these cool tiles but they are treated as similar to all other tiles. The tile power for such cases is low but the rest of the chip power used as temperature predictor becomes high and thus overestimates the temperature. This error can be easily rectified by adding one more step of checking the very low power tiles with low power 2D and 3D neighbors in a given design before feeding the predictors into the model. For these tiles, the lumped power value of rest of the chip can be scaled down before analyzing the temperature.

2.4.3 Runtime comparison

Since the developed model is a compact model with a simple mathematical relation obtained by regression, it is many orders of magnitude faster than full GDS-level analysis and

compact resistive network analysis methods. This very important property helps in direct temperature estimation during a larger part of the design process. Table 6 summarizes the runtime comparison with GDS-level FEA simulation and Hotspot [26]. The runtime is reported after the analysis of a 3-tier 3D design with 1.3mm x 1.3mm footprint. Hotspot is run for steady-state thermal analysis with 3D stacking using a 16x16 grid network such that each grid's size is $81.25\mu m$ X $81.25\mu m$, that is similar to the tile size used in the presented model. The new model is 4.9 x 10^6 times faster than FEA simulation and 2.6 x 10^4 faster than hotspot analysis for 3D stacking.

2.5 Thermal-aware Floorplanning

2.5.1 Floorplanning algorithm

Simulated annealing of sequence pair representation of floorplan is used to obtain the best floorplan depending on the weighted cost function specified. The non thermal-aware floorplanner excludes the maximum temperature of chip from the cost function. Since this is a monolithic design, 3D via does not occupy any area and hence, the number of 3D connections is not included in the cost function. It is known that larger area will help in reducing temperature by reducing the power density. However, area is directly proportional to cost, especially in miniaturized systems. Therefore, the floorplanner is tuned to start optimizing temperature only after the specified area constraint is satisfied. Also, there is a trade-off between maximum temperature reduction and total wirelength to have minimum performance overhead. More wirelength will increase total net switching power in the final design which may increase temperature further. However, if the blocks are not given freedom of movement within the constrained area, the solution space for temperature optimized floorplans within that area becomes smaller and there won't be significant temperature reduction. This freedom of movement of blocks implies wirelength overhead in the overall floorplan. Therefore a step by step process is used to obtain the temperature optimized floorplan.

First, the non-thermal floorplanner is run without any temperature cost to obtain the

expected wirelength value. In the next step, given a certain slack on this wirelength, wirelength and maximum temperature is included in the initial cost function. Once, the wirelength goal is met, only temperature within that area and wirelength constraint is minimized. Any floorplan solution which violates the area and wirelength requirement is rejected. The floorplanner is also run with only 5% area slack to give more room for improvement. The final result obtained can be below this area limit. The fact that the developed thermal model is extremely fast with good accuracy enables faster temperature profile evaluation of every sequence pair without any runtime issues and hence minimizes the maximum temperature.

For a design with *B* blocks, *N* nets and *T* thermal tiles, the complexity changes from *O(BlogB+N)* to *O(BlogB+N+T)* by including temperature evaluation [53]. The wirelength calculation for all nets for a given sequence pair is the major time complexity in the floorplanning process. Therefore, the addition of thermal analysis with the developed model, which uses $100\mu m \times 100\mu m$ tiles, has insignificant overhead even with millions of moves during simulated annealing.

After obtaining the temperature optimized floorplan, place and route of the design is carried out using Encounter and then the power and timing analysis is carried out to verify that there is no performance overhead. All benchmarks were designed to meet the specified timing requirement with minimal change in worst negative slack. A final full GDS-level thermal FEA was carried out with the specific package structure to check the maximum temperature.

2.5.2 Floorplanning results for conventional package

Two benchmark circuits are reported for floorplaning comparison. The FFT benchmark is obtained at RTL level from Opencores and has 49 blocks of different sizes with 1400 inter-block nets. The industry circuit benchmark was obtained directly at block level only with inter-block nets and block powers. It has 32 blocks with 9203 nets. Since the verilog netlist

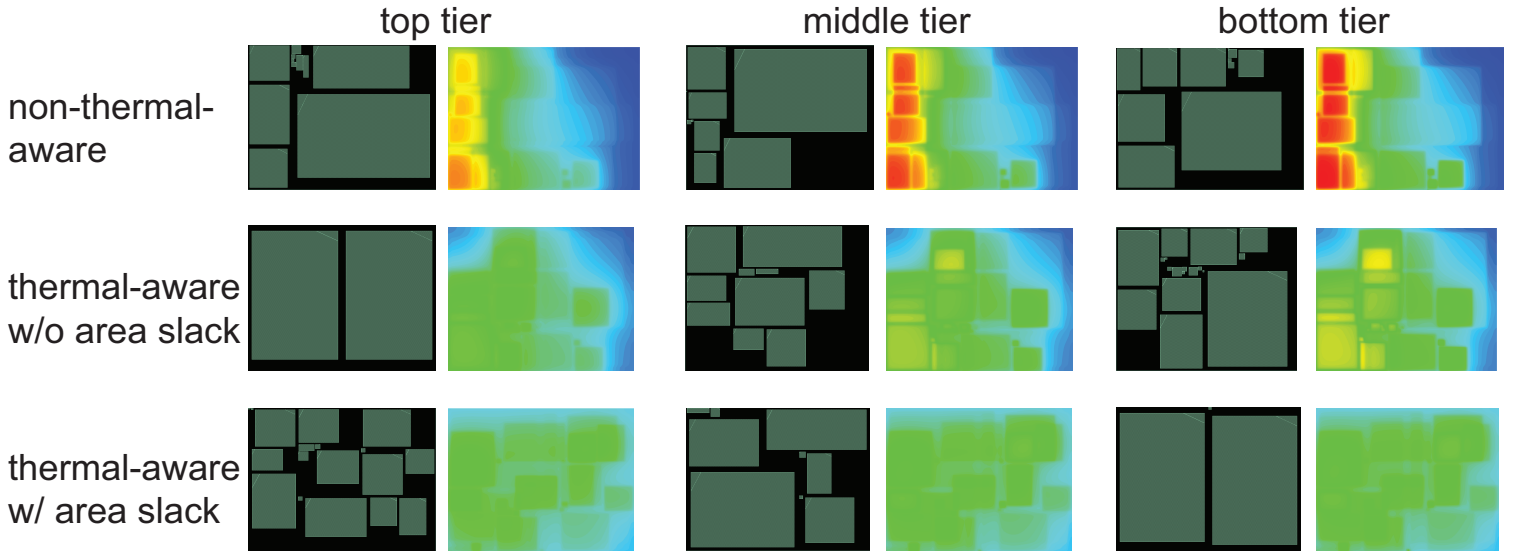


Figure 10: 3-tier floorplanning layouts (ind_ckt benchmark with conventional package structure) with corresponding absolute temperature maps. The temperature range is [47°C, 68°C].

Table 7: Thermal-aware floorplanning with temperature model developed for conventional package structure

	Footprint ($\mu m x \mu m$)	Si Area (mm^2)	Inter-block WL (m)	Max Temp above room(^oC)	Avg Temp above room(^oC)	Temp Gradient(^oC)	Floorplan Runtime(sec)
cf_fft_256_8							
2D	1181 x 1147	1.36	0.56	22.12	13.57	10.39	-
Non-thermal	745 x 939	1.40 **(1.00)**	0.34	33.38 **(1.00)**	26.26	10.19	1452 **(1.00)**
2-tier Thermal (w/o area slack)	762 x 920	1.40 **(1.00)**	0.45	31.62 **(0.94)**	25.88	8.37	1723 **(1.18)**
Thermal (w/ area slack)	867 x 849	1.47 **(1.05)**	0.45	27.36 **(0.82)**	24.37	5.56	1780 **(1.23)**
Non-thermal	580 x 824	1.43 **(1.00)**	0.34	48.05 **(1.00)**	39.14	13.00	1486 **(1.00)**
3-tier Thermal (w/o area slack)	577 x 829	1.43 **(1.00)**	0.37	44.20 **(0.92)**	38.47	9.26	1769 **(1.19)**
Thermal (w/ area slack)	891 x 560	1.50 **(1.05)**	0.35	42.84 **(0.89)**	36.69	11.31	1808 **(1.22)**
ind_ckt							
2D	3939 x 3525	13.89	10.18	15.02	10.71	6.82	-
Non-thermal	3680 x 1994	14.68 **(1.00)**	6.43	26.57 **(1.00)**	19.76	11.19	3228 **(1.00)**
2-tier Thermal (w/o area slack)	3603 x 1994	14.37 **(0.98)**	6.86	25.20 **(0.95)**	19.74	9.99	5552 **(1.72)**
Thermal (w/ area slack)	3050 x 2491	15.19 **(1.03)**	7.33	23.89 **(0.90)**	18.93	9.44	5677 **(1.76)**
Non-thermal	2591 x 1960	15.24 **(1.00)**	5.54	40.89 **(1.00)**	28.66	20.30	3600 **(1.00)**
3-tier Thermal (w/o area slack)	2452 x 2070	15.22 **(1.00)**	5.91	35.73 **(0.87)**	28.72	13.80	6471 **(1.80)**
Thermal (w/ area slack)	2454 x 2037	15.00 **(0.98)**	6.29	32.03 **(0.78)**	28.41	8.00	6074 **(1.69)**

of the industry circuit and the intra block information is not provided, there is no place and route step of this the design and only the HPWL is reported. The block power numbers result in a large temperature gradient in the non thermal aware design and the inclusion of temperature cost evaluated using the thermal model improves the temperature profile significantly. The inter-block nets' switching power is obtained by timing and power analysis using Synopsys PrimeTime and is considered in the final GDS-level thermal analysis. The purpose is to ensure that even with slight power increase due to increased wirelength, the thermal aware floorplan results in reduced temperature. Since inter block wirelength is very less compared to total wirelength, there is negligible increase in interconnect power due to increase in inter-block wirelength.

The results of the different cases implemented during floorplanning with conventional package are summarized in Table 7. The implementations for 2-tier and 3-tier 3D designs are presented for conventional package structure. 2D design metrics are also given for reference. Since the runtime is dependent on number of blocks, number of nets (for wirelength calculation) and size of the chip (for temperature estimation), there is different runtime for the different designs, but the increase in runtime due to thermal analysis is well within tolerable limits.

There is significant reduction in maximum temperature given the fact that there is minimum area overhead, therefore satisfying the purpose of the thermal-aware floorplanning (Figure 10). The developed thermal-aware floorplanner tries to reduce the gradient of the temperature variation as the average power density of the chip will remain the same because of the same chip area with the same total power dissipation. The floorplanning process avoids stacking of high power density blocks and also forces such blocks to tiers which are closer to the heat sink. The larger and low power density blocks are placed in the critical tiers. The 3-tier designs show more degree of improvement because of more options to move the blocks around. All of this becomes feasible only because of the fast and accurate monolithic 3D IC temperature estimation model.

Table 8: Comparison with 3DFP [27] (FFT benchmark)

	Footprint ($\mu m x \mu m$)	Si Area (mm^2)	Inter-block WL (m)	Max Temp above room(^oC)
2-tier				
3DFP [27]	1005 x 735	1.48	0.60	27.63
New FP	867 x 849	1.47	0.45	27.36
3-tier				
3DFP [27]	972 x 518	1.51	0.46	45.81
New FP	891 x 560	1.50	0.35	42.84

2.5.3 Comparison with state-of-the-art

Cong *et al.* show 56% temperature reduction but a 21% area increase which is significantly high overhead [17]. A fast but less accurate hybrid resistive model and another accurate but relatively slow resistive model is used selectively within the floorplanning process.

Power density and total 3D overlap in the cost function to incorporate thermal awareness during design has been used for thermal aware 3D floorplanning [27]. The tool called 3DFP is available for public use. The new thermal model is more effective than 3DFP in the design process because it directly gives an accurate measure of temperature. 3DFP is used on the same benchmarks and the results are compared. Since, the number of moves during annealing and other annealing parameters differ in the two floorplanners, only the quality of the floorplan results is compared, and not the total runtime.

Table 8 shows the comparison results of 3DFP and the developed thermal-aware floorplanner for the FFT benchmark. With the help of direct temperature measurement during annealing using the fast and accurate model, better floorplans are designed in all respects viz. area, wirelength and temperature.

2.5.4 Thermal floorplanning for modern mobile package

Table 9 shows the results for thermal aware floorplanning for 3-tier 3D ICs with mobile package structure for the two benchmarks. The power densities have been scaled for the designs to satisfy the peak temperature limitations for mobile package structures. Figure 11

Table 9: Thermal-Aware floorplanning with temperature model developed for modern mobile package (no heat sink)

	Footprint ($\mu m x \mu m$)	Si Area (mm^2)	Inter-block WL (m)	Max Temp above room(^oC)	Average Temp above room(^oC)	Temp Gradient(^oC)	Floorplan Runtime(sec)
cf_fft_256_8							
2D	1181 x 1147	1.36	0.56	11.96	10.91	1.88	-
Non-thermal	580 x 824	1.43 **(1.00)**	0.3	32.78 **(1.00)**	29.13	5.10	1486 **(1.00)**
3-tier Thermal (w/o area slack)	552 x 853	1.41 **(0.99)**	0.32	31.42 **(0.96)**	28.77	4.21	1985 **(1.34)**
Thermal (w/ area slack)	667 x 739	1.48 **(1.04)**	0.34	28.98 **(0.88)**	28.23	1.43	1889 **(1.27)**
ind_ckt							
2D	3939 x 3525	13.89	10.18	23.24	16.43	12.00	-
Non-thermal	2591 x 1960	15.24 **(1.00)**	5.54	39.81 **(1.00)**	27.07	23.06	3600 **(1.00)**
3-tier Thermal (w/o area slack)	2420 x 2097	15.22 **(1.00)**	5.97	38.79 **(0.97)**	27.15	21.06	6564 **(1.82)**
Thermal (w/ area slack)	2701 x 1949	15.79 **(1.04)**	6.27	35.18 **(0.88)**	23.07	20.23	6962 **(1.93)**

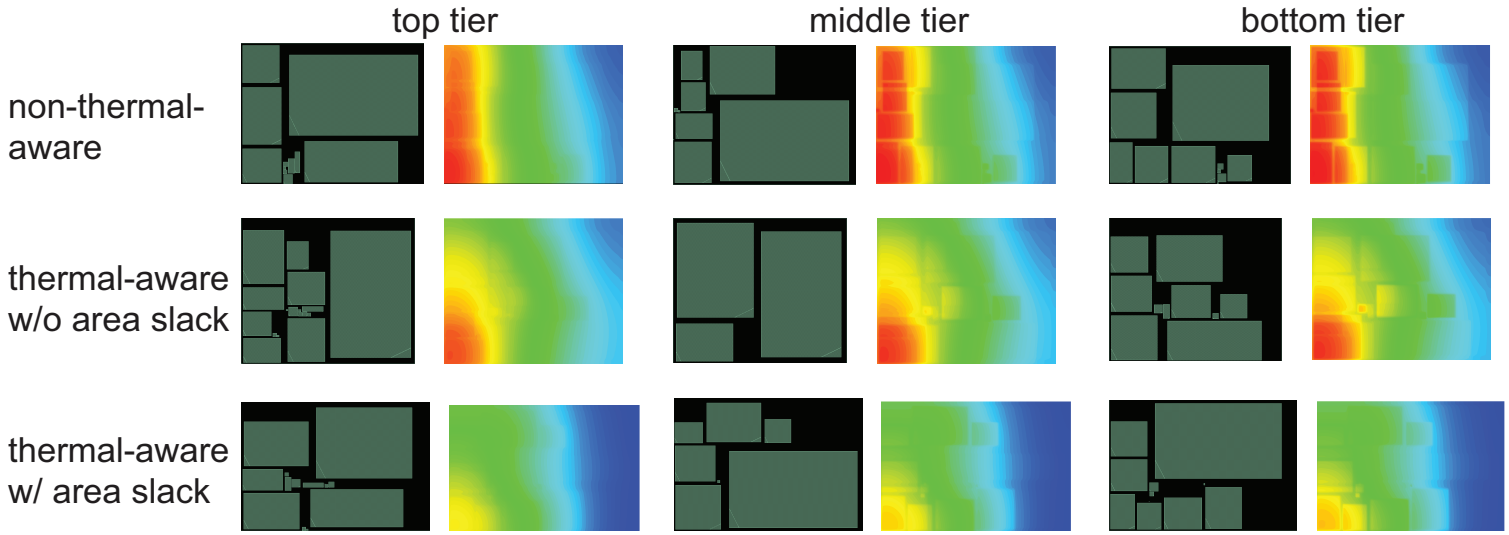


Figure 11: 3-tier floorplanning layouts (ind_ckt benchmark with mobile package structure) with corresponding absolute temperature maps. The thermal-aware floorplans avoid stacking of high power density blocks and keep low power density blocks in middle tier. The temperature range is [42°C, 67°C].

shows the floorplanning results along with their temperature maps for industrial circuit benchmark with mobile packaging structure. For such type of packaging, the middle tier is most critical as heat dissipates in both directions. The thermal model correctly maps the mobile package system along this line and the larger, low power density blocks get placed in the middle tier (Tier1) without any area overhead. For the thermal-aware floorplanning with no area overhead, Tier1 ends up with only three large blocks with low power density reducing maximum temperature.

2.6 Mobile Package Optimization: Impact of the Materials

The different layers which play an important role in thermal behavior of mobile packages due to two-way heat flow were presented in Section 2.3. In this section, the impact of thickness and conductivity of some of these materials in mobile package on the maximum temperature are studied. In particular, the difference in impact on 2D ICs and 3D ICs is examined and the fact that a good change in package properties is more beneficial for 3D ICs than 2D ICs is established. The different material properties varied are the ones specified with range of values in Table 2. Figure 12 plots the maximum temperature of 2D IC and 2-tier 3D IC with change in various material thicknesses and conductivities. 3D ICs have multiple layers of power dissipation source while 2D ICs have just one device layer dissipating power. Therefore, the impact of changing the thickness and conductivities of package layers is more prominent in 3D IC than 2D IC.

Thermal Interface Material (TIM) is a necessary layer to have smooth continuous contact between the uneven bulk surface and the EMI (or heat spreader for conventional packages). They are poor conductors of heat (< 5 W/mK) but provide a better and continuous thermal interface than silicon-air and air-metal interface. Because the TIM provides a high resistive path to heat flow, changes in the thickness of the layers beyond TIM (EMI and graphite) have negligible impact. The thermal circuit is equivalent to a large resistance (TIM here) in series with a small resistance (EMI) whose value changes with change in

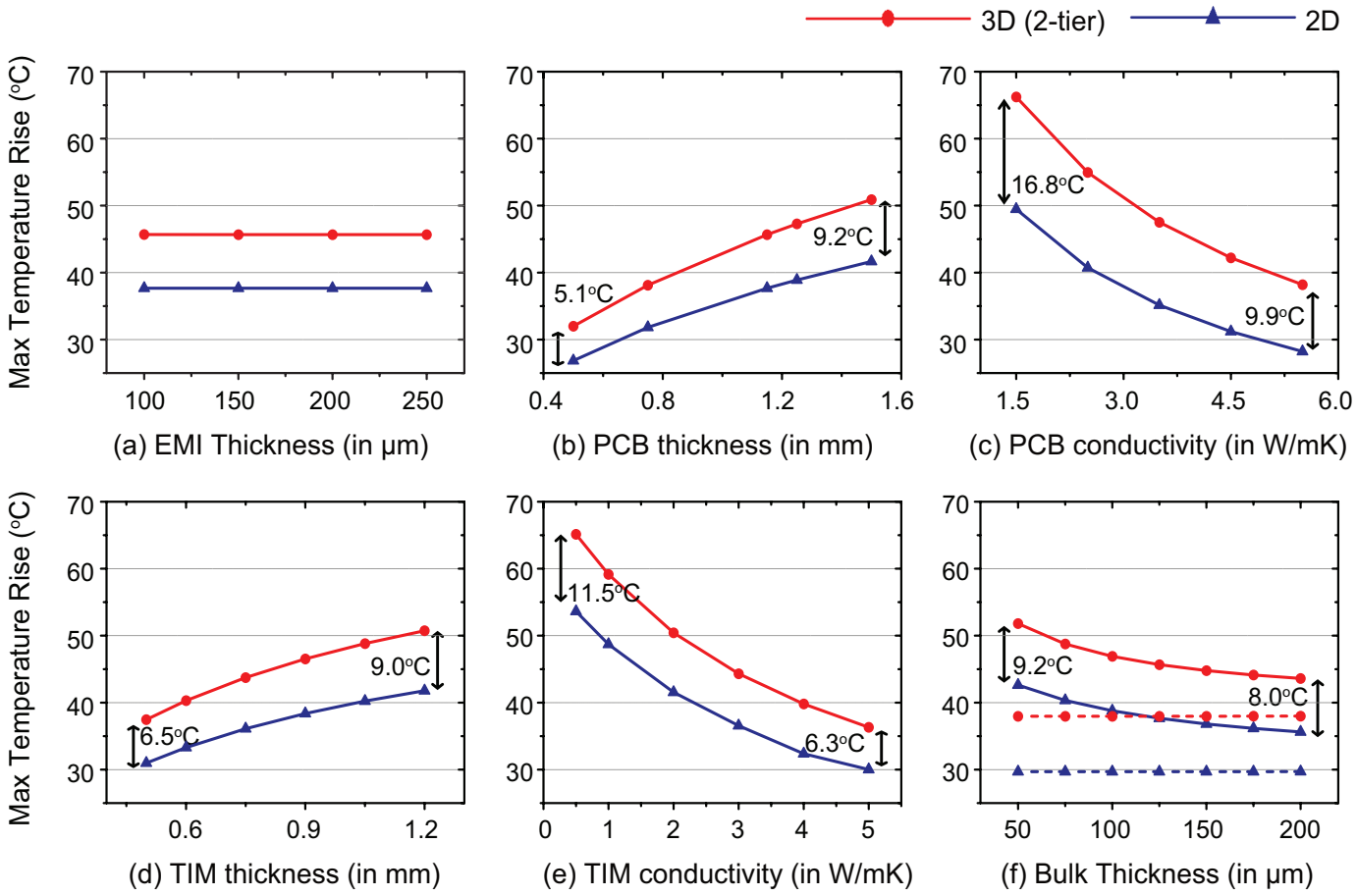


Figure 12: Impact of change of various material thicknesses and conductivity on maximum temperature of ind_ckt benchmark with mobile package structure for 2D IC and 3D IC (2-tier). Dotted lines in (f) is the average temperature variation.

thickness but has minor impact on equivalent resistance. This is shown in Figure 12a where change in EMI thickness has no impact on the maximum temperature of 2D IC as well as 3D IC.

Figure 12b and Figure 12c show the impact of change of PCB thickness and vertical conductivity respectively. Since the PCB is the major heat flow path in the downward direction (Figure 5b), improvement in its thermal resistance reduces maximum temperature significantly. Reduction in thickness or increase in vertical conductivity both contribute to reduced thermal resistance. The change is more prominent in 3D ICs because the bottom tier now finds a much lower resistive path in the downward direction and the amount of heat transferred towards the upper tier is reduced therefore reducing the degree of thermal coupling in the two tiers.

TIM has a similar impact as PCB as shown in Figure 12d and Figure 12e. The vertical conductivity change brings about a larger degree of maximum temperature change because the conductivity value itself is changed from 0.5-5 W/cm^2 which is a 10X increase. Though not shown in the plots, the average temperature also follows the same trend as the maximum from Figure 12a-12e as most of these layers are towards the end of the equivalent thermal circuit. Handle bulk on the other hand is an intermediate layer in the heat flow path. Silicon being a reasonably good conductor helps more in lateral spreading to reduce the temperature gradient across the design but has no impact on the overall average since there is a poor conducting TIM layer further up in the path. Figure 12f shows the change in maximum temperature with change in handle bulk thickness. The average temperature is also shown in dotted lines and has no change with change in bulk thickness but the maximum temperature reduces due to change in gradient. Also the difference in impact on 2D IC and 3D IC is not very high as observed for the PCB and TIM layers.

Therefore, the package structure plays an important role in determining the maximum temperature and the same change in package properties exhibit more benefits for 3D ICs with mobile packages. This can be used to plan a good package structure to start with or as

a post physical design technique to further improve the thermal reliability of 3D ICs after obtaining a thermal aware layout.

2.7 Summary

This chapter presented the unique thermal properties of monolithic 3D ICs and compared their thermal behavior with TSV-based 3D ICs. It was observed that due to the absence of bulk silicon substrate in monolithic technology, their is no lateral spreading near the device layer. Also the very thin ILD and absence of bonding layer results in heavy vertical thermal coupling and improves the temperature profile of the tiers away from the heat sink compared to TSV-based 3D ICs.

The properties were utilized to develop a methodology to obtain package-aware fast and accurate thermal analysis model for monolithic 3D ICs with different number of stacking layers using non-linear regression. These models were verified against full chip FEA thermal simulations and found to be highly accurate.

Next, the models were used in a thermal-aware floorplanner to obtain significant temperature reduction with minimum or no area overhead for both conventional packages with heat sink and mobile packages. The impact of material property changes in mobile package structure to enable thermal package optimization for 3D ICs was also studied.

CHAPTER III

POWER DELIVERY IN MONOLITHIC 3D ICS

Power delivery network (PDN) is an integral part of any circuit design. In conventional 2D ICs, few top metal layers are dedicated for PDN, while most of the signal routing happens in the bottom and intermediate layers. Therefore, there is no conflict of resource usage between signal and power routing. However, for 3D ICs, PDN and signal have to pass through top metal layers of one tier to connect to cells in the other tiers. This affects the resource allocation for signal routing and worsens congestion. PDN tradeoffs among wirelength, power, and thermal are more pronounced in monolithic 3D ICs than TSV-based 3D and 2D designs because of the higher integration density which uses many MIVs leading to severe competition between signal and power connections. The relative impact worsens at advanced technology nodes due to higher congestion of interconnects. The increase in signal wirelength leads to higher switching power dissipation, which significantly contributes to total power and worsens thermal behavior.

In this chapter, a comprehensive study on the impact of power delivery network (PDN) on full-chip wirelength, routability, power, and thermal effects in gate-level monolithic 3D ICs across different technology nodes is presented. The new challenges in PDN for M3D are identified and their impact quantified. Various PDN design optimization techniques for monolithic 3D ICs at different nodes are developed and tested with full chip designs to reduce PDN impact on signal wirelength and total power under the given IR drop budget.

3.1 Motivation and Background

The side-view of a typical two-tier monolithic sequential structure with seven metal layers in each tier is shown in Figure 13. The orientation of this figure is as per the fabrication sequence and not in flip-chip configuration. The device layer thickness is around 30nm

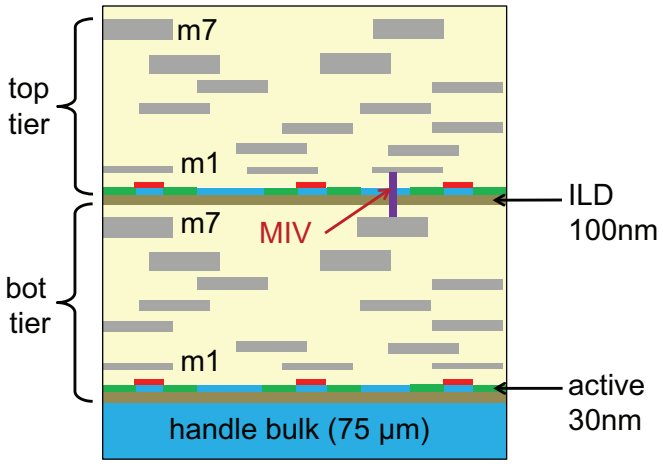


Figure 13: Side-view of 2-tier monolithic 3D IC structure (with seven metal layers in each tier)

and the inter tier dielectric (ILD) which separates different tiers is about 100nm thick. The monolithic inter-tier via (MIV) diameter of <100nm allows high integration density with significant wirelength reduction in M3D. The MIV connects the top metal of bottom tier to the bottom metal of the top tier.

Since, gate-level monolithic 3D ICs provide ultra high integration density by the stacking of dies, power density of the chip and also the current demand per unit area is increased. This in turn complicates the the power delivery network design. Interconnects do not scale at the same rate as devices and the parasitic resistance of wires is higher at advance nodes [10, 11]. Therefore, interconnect impact worsens as devices become smaller. Monolithic 3D ICs helps in reducing the interconnect length significantly by allowing tens of thousands of 3D MIVs. However, the presence of power delivery network (PDN) reduces the routing resources and adversely affects the total signal wirelength. This impact is more severe in M3D due to heavy usage of top metals for 3D routing along with PDN routing. PDN is always important in any design, but due to sequential 3D layers with very high integration density in M3D, optimal PDN planning becomes much more critical.

Prior works focusing on power reduction benefits in gate-level monolithic 3D ICs and related CAD methodologies [45, 12] do not address any thermal issues and totally ignore

power delivery design and challenges. Wei *et al.* have shown that PDN helps in reducing the temperature in monolithic 3D ICs with an example of OpenSPARC T2 processor core [54]. Though the presence of PDN helps in improving the thermal conductivity and reduces maximum temperature, the same power dissipation of the blocks with or without PDN is assumed, ignoring its impact on increased congestion during signal routing. This congestion results in increased signal wirelength and hence increased net power, especially in advanced technology nodes. Also, the power simulation has been carried out at the architectural level and does not include parasitic impacts from full layout-extraction. Billiont *et al.* further studied the impact of tungsten in bottom tier along with the presence of PDN in designs but the primary focus was development of CAD flow [9]. Panth *et al.* developed a tier-partitioning technique to handle power-delivery and thermal trade-off in M3D for mobile applications [44].

Some other 3D PDN works focus only on TSV-based 3D ICs or 3D PDN simulation and analysis techniques. A 3D IC floorplan and power-ground co-synthesis tool is developed in [23]. However, only block level floorplanning and power/ground design is considered for TSV-based 3D designs. There is no discussion on the PDN inside the blocks. The total intra-block wirelength heavily dominates the total wirelength in any design. In other works, Luo *et al.* developed 3D IC power delivery networks benchmark for research purposes [39]. Various sizes of 3D designs are covered, but all of them are TSV-based and at the block level.

Full chip impact study of PDN design on monolithic 3D ICs has not been studied much. With advancement of technology and scaling limits, gate-level monolithic 3D ICs enable further extension with significant power and scaling benefits. Therefore, it is important to study and understand all factors influencing this technology, and develop methods to maximize the overall benefits, especially in advanced nodes. PDN optimization is one of such important factors.

3.2 Design and Analysis Setup

Two benchmark circuits at three different technology nodes are designed and analyze to understand the PDN impact on physical design for 2D and monolithic 3D. The benchmarks used are (1) OpenSPARC T2 single core [43] and Advanced Encryption Standard (AES) circuit [42]. 28nm, 14nm and predictive 7nm technology library are used for the design implementations and analysis.

The 2D physical design is carried out using standard RTL-GDSII flow in Cadence Encounter. The 3D designs are first placed using the same tool in a 2D fashion but with doubled capacity and then partitioned by placement driven partitioning [47]. Since, the focus of this study is the impact of PDN on the full design, any good partitioning methodology [45, 9, 47] will suffice and the relative impact of similar PDN on signal wirelength will be similar. MIV planning and insertion is carried out by stacking the individual tiers together into a single 3D metal stack with appropriate cell-pin locations and then using Encounter nanoroute tool to get the via locations ([45, 9, 47]). After 3D Placement and MIV planning, tier by tier routing is performed to get the final designs. Two types of 2D and 3D designs are implemented respectively, with similar placement of cells. One is routed without any power or ground wires in the metal layers, while the other one has power/ground wires in the relevant metal layers. The circuits are designed to meet similar timing constraints in both 2D and 3D at the respective nodes and are therefore iso-performance designs. Power analysis is carried out on these iso-performance designs using Synopsys PrimeTime. Figure 14 shows example layouts of the placement and routing of both the tiers of T2 benchmark with 28nm technology. Power/ground wires are not shown for clear view of signal routing.

3.2.1 Technology scaling

Technology node is a major aspect that will affect the power benefits of M3D and the impact of PDN. As per recent ITRS roadmap [28], interconnect scaling lags behind device

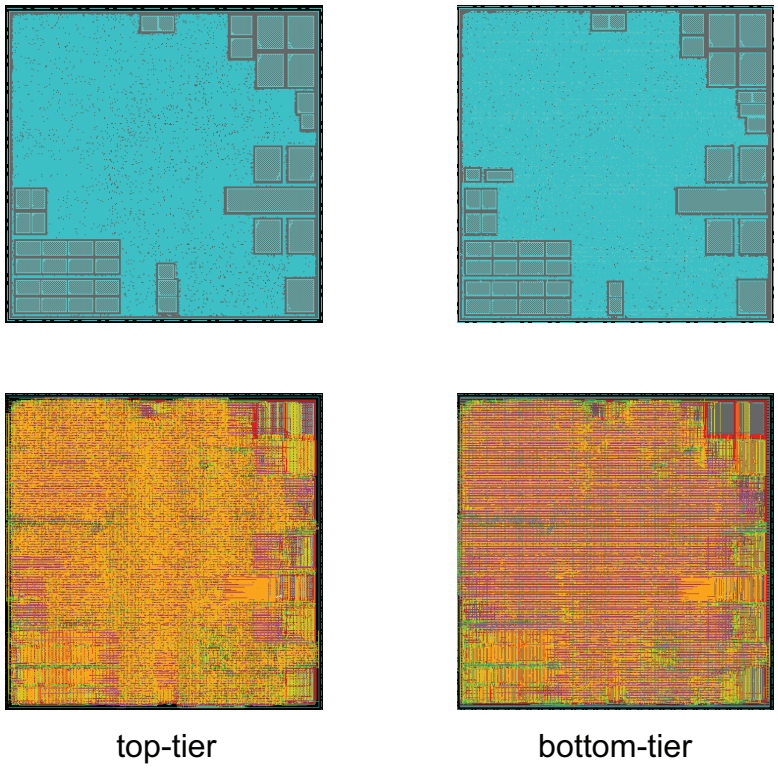


Figure 14: OpenSPARC T2 layouts in monolithic 3D. The top row is the placement/floorplan in each tier and the bottom row shows the overall signal net routing. PDN is not shown in routing for clarity.

scaling. Resistivity of copper wires increase rapidly at advanced nodes due to increased electron scattering at the grain boundaries and surfaces. At 7nm, copper has 3.7X larger resistivity than 45nm node. Therefore, it is imperative to study and analyze the impact of PDN on M3D at future nodes as well.

Full process design kits (PDK) for 45nm and 28nm are openly available for use in research. However, the new open-source 15nm PDK [40] is not fully developed yet and lacks important information for full RTL-GDSII designs. In addition, there is no real 7nm PDK available in any form. Therefore, for this study, predictive libraries were built at 14nm and 7nm nodes for relative comparison based on Synopsys 28/32nm PDK [52] and ITRS data (advancement of two and four nodes respectively relative to 28nm node). This approach is motivated from earlier works ([11, 14]) that used ITRS based scaled PDKs for

Table 10: Power/delay comparison of basic cells (X1 size) at different nodes. (input transition time = 32ps, load Cap = 1fF)

		Cell Power (fJ)		Rise Delay (ps)		Fall Delay (ps)	
INV	28nm	0.350	(1.00)	5.79	(1.00)	6.01	(1.00)
	14nm	0.147	(0.42)	5.05	(0.87)	5.29	(0.88)
	7nm	0.054	(0.15)	4.63	(0.80)	4.71	(0.78)
NAND2	28nm	0.559	(1.00)	20.59	(1.00)	18.06	(1.00)
	14nm	0.252	(0.45)	17.12	(0.83)	15.89	(0.88)
	7nm	0.099	(0.18)	15.47	(0.75)	14.44	(0.80)
NOR2	28nm	0.606	(1.00)	21.69	(1.00)	23.35	(1.00)
	14nm	0.270	(0.45)	18.28	(0.84)	20.05	(0.86)
	7nm	0.117	(0.19)	15.35	(0.71)	17.68	(0.76)

research.

The cell layouts are scaled as per ideal area-scaling ratio of 50% per technology node. Though this is optimistic scaling, it allows a direct relative comparison of PDN impact on signal routing and hence wire power across different technologies. Therefore, the same cell in 28nm technology is 4X the size of 14nm cell and 16X the size of 7nm cell in terms of area. All cells are 11-track layouts i.e. the height of the cell is 11X that of minimum metal1 pitch (as provided in 28nm PDK). Transistor models are used from ASU-PTM [48] to obtain the new timing libraries for the 14nm and 7nm standard cells. The nominal voltages for 28nm, 14nm and 7nm PDKs are 1.05V, 0.8V and 0.7V respectively. While 1.05V is as provided in the 28nm PDK, the other values are based on the transistor model data. The high input gate-capacitance and internal power of cells (relative to the expected reduction with scaling) of finFETs used in post-16nm nodes are accounted for in the 14nm and 7nm cell libraries.

Table 10 shows the power-delay comparison of some basic cells in the three technology nodes. The libraries built for 14nm and 7nm closely follow the predictive technology trend in [10, 11], where 45nm library has been used as the reference. In this chapter of the work, 28nm library [52] is used as the reference. For the new libraries, new interconnect technology (tch) filesare generated with extensive electromagnetic (EM) simulation using Cadence Techgen for use in accurate full chip layout parasitic extraction. RC parasitic

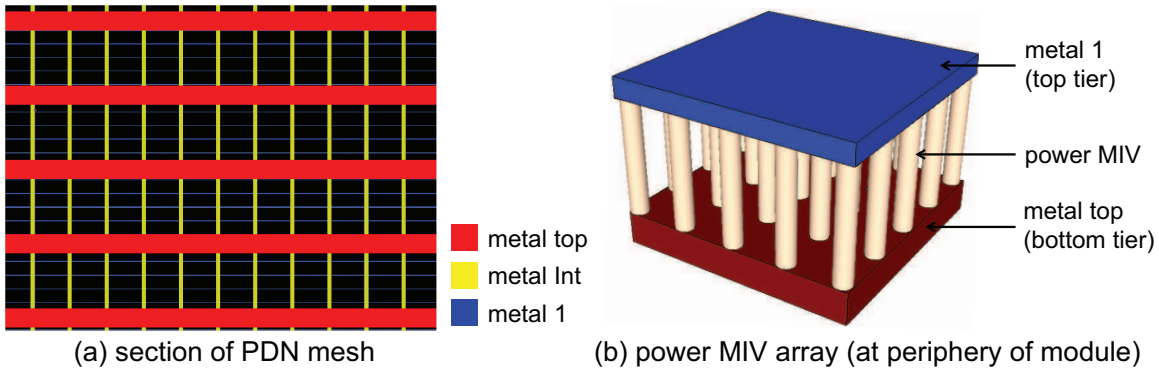
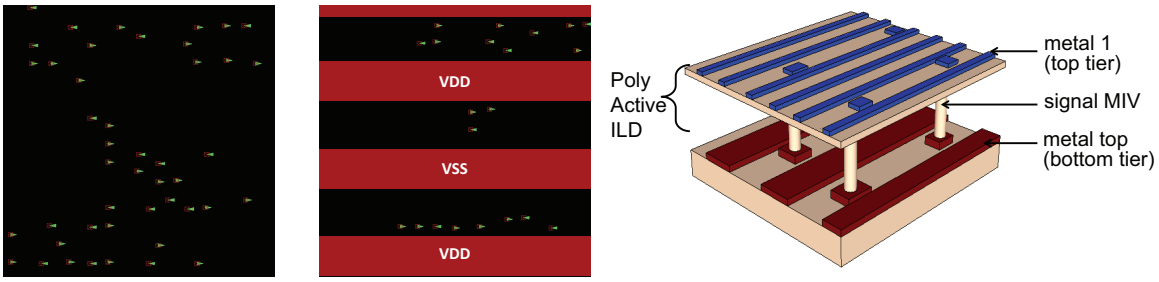


Figure 15: (a) Section of power/ground mesh structure, (b) a power MIV array at periphery.

values and material properties are based on ITRS projections [28]. While capacitance per unit length remains almost the same across nodes (~$0.2fF/\mu m$) due to use of low-K dielectrics, the metal layers become severely resistive at advanced nodes due to the larger copper resistivity and the smaller metal width/thickness. But the length of nets also reduce, so the overall impact cannot be directly predicted without full chip design and analysis.

3.2.2 Power delivery network designs

In this work, standard power delivery network design methodology is used, which uses topmost metal layer for global wires and then intermediate metal layers to connect to the metal1 VDD and VSS rails [44]. Figure 15a shows the PDN mesh layout structure. The density of power/ground wires is determined such that the maximum IR drop in 2D designs is limited to 5% of the supply voltage (VDD). Similar PDN layout is maintained across all technologies with respective pitches/widths of metals and supply bumps. The same pitch of PDN wires is used in 2D and 3D designs for the same technology and then full chip impact comparisons are carried out. The benchmarks are parts of larger SoC designs and hence sub-modules or cores of a bigger design (e.g. T2 single core is one unit of 8-core T2). The power supply to such designs is delivered mostly through rings and stripes that are connected to the full SoC's PDN. These supplies are then distributed using the intra-module PDN. The C4 bumps in flip-chip designs supply power to the full SoC which in



(a) without any PDN (b) with VDD and VSS stripes (c) signal MIV landing pad locations constrained

Figure 16: Impact of PDN on MIV landing pads a) MIVs freely distributed without any PDN blockages in top metal b) PDN blockages affect MIVs in top metal c) isometric view showing the constraints on signal MIV landing pad locations in top metal and metal1 of the next tier.

turn distributes it to these different modules through their rings. Therefore, in this analysis, input power/ground supply is only provided at the periphery of the chips. This same relative density of PDN usage is maintained in both 2D and M3D designs. Therefore, it gives a direct comparison on PDN impact. Addition of more power/ground inputs with same mesh will only improve the IR drop numbers in both 2D and 3D designs.

OpenSPARC T2 single core contains register-file modules which block up to metal4. Therefore, seven metal layers for full signal routing in T2. AES is a gate only design and relatively smaller and five metal layers are sufficient for full routing. These numbers are determined after performing experiments to verify the total routing requirement.

For T2 single core benchmark, metal7 is used for horizontal wires and metal4 for the intermediate vertical wires in the power/ground mesh. The preferred orientations are used as specified in the technology. 40% of the top metal and 20% of the intermediate metal is used for PDN. The top metal power/ground wires are made wider while the intermediate metal4 wires are relatively less wide. The frequency of occurrence of the wires are determined by the percentage usage as mentioned earlier and all the wires are placed equidistant from each other. The VDD and VSS cell rails run on metal1 cell rows as usual. For the AES design where five metal layers are used, metal5 is used for the horizontal power/ground wires. All the designs are circumscribed with VSS and VDD rings.

For 3D designs, 3D power ground connections are in the form of MIV arrays distributed

along the periphery (Figure 15b). Since a single MIV is extremely thin, it offers very high resistance to the supply paths (~10Ω). Therefore, an array of such MIVs is used for a single supply connection. 15 x 15 arrays ($3\mu m$ x $3\mu m$ at 28nm node) are used here. Such arrays are not placed inside the main area of the design because it consumes and blocks extra silicon area. Also reduced footprint in 3D designs helps reduce path resistance from periphery to center.

Full IR drop analysis is performed using Cadence VoltageStorm. 3D technology files are generated to allow the tool to carry out 3D IR drop analysis. The 3D technology file has the basic stack-up information of the two tiers of monolithic 3D IC along with the vertical dimensions of the ILD and MIVs connecting the tiers . Current sources (cell locations) are placed at the respective metal1 rails of each tier and the full two-tier IR drop analysis is carried out. The values of the current sources are determined using the detailed cell-level power analysis with PrimeTime. For 3D designs, it is always the tier farther from the external supplies that is affected most. In the study, the bottom tier of Figure 13 shows maximum IR drop since it is away from the external power supply input which is on the top metal of top tier.

3.3 New PDN issues in Monolithic 3D ICs

Power Delivery Network (PDN) with low max IR drop is an integral part of any digital circuit. In general, the top metal layer(s) is completely dedicated for power delivery depending on the number of supply voltages required for the design. Instead of having a long via-stack to metal1 VDD and VSS rails which connect directly to the standard cells, one or more intermediate metal layers are also used in between to redistribute the supply and reduce the resistive drop [44]. Because of the usage of these metals for PDN, there are less routing resources available for signal nets. This results in many detours for signal nets and hence increased wirelength. This increased wirelength and coupling results in more wire power dissipation.

Table 11: Detailed comparison of impact of power delivery network (PDN) on 2D IC and monolithic 3D IC designs of OpenSPARC T2 -Single core benchmark for different technology nodes. All $\Delta\%$ numbers are evaluated relative to the respective w/o PDN design.

Design Style		Footprint (μm x μm)	# MIVs	Wirelength WL (m)	$\Delta\%$	Wire Power (mW)	$\Delta\%$	Cell-Pin Power (mW)	Cell-Internal Power (mW)	Total Power (mW)	$\Delta\%$
28nm node (0.667GHz)											
2D	w/o PDN	1600 x 1600	-	17.89		273.2		185.2	169.2	754.3	
	w/ PDN			19.12	**+6.9%**	293.4	**+7.4%**	185.5	173.9	779.4	**+3.3%**
3D	w/o PDN	1140 x 1140	140,546	15.33		212.6		157.9	159.0	642.9	
	w/ PDN		116,727	17.09	**+11.5%**	242.0	**+13.8%**	158.1	162.7	676.1	**+5.2%**
14nm node (0.770GHz)											
2D	w/o PDN	800 x 800	-	9.16		137.7		101.0	90.3	405.0	
	w/ PDN			10.13	**+10.6%**	162.5	**+18.0%**	100.6	95.9	435.0	**+7.4%**
3D	w/o PDN	570 x 570	120,099	8.32		121.5		86.2	84.5	360.2	
	w/ PDN		107,371	9.91	**+19.1%**	161.5	**+32.9%**	85.8	90.8	406.1	**+12.7%**
7nm node (1.000GHz)											
2D	w/o PDN	400 x 400	-	4.59		108.3		57.5	55.2	268.7	
	w/ PDN			5.26	**+14.6%**	137.7	**+27.1%**	57.5	55.2	301.1	**+12.0%**
3D	w/o PDN	285 x 285	121,022	4.16		96.2		49.2	49.3	240.0	
	w/ PDN		108,068	5.17	**+24.3%**	137.4	**+42.8%**	49.9	52.8	285.4	**+18.9%**

Table 12: Detailed comparison of impact of Power Delivery Network (PDN) on 2D IC and Monolithic 3D IC designs of AES benchmark for different technology nodes. All $\Delta\%$ numbers are evaluated relative to the respective w/o PDN design.

Design Style		Footprint (μm x μm)	# MIVs	Wirelength WL (m)	$\Delta\%$	Wire Power (mW)	$\Delta\%$	Cell-Pin Power (mW)	Cell-Internal Power (mW)	Total Power (mW)	$\Delta\%$
28nm node (1.250GHz)											
2D	w/o PDN	750 x 750	-	3.64		92.3		77.4	63.6	247.0	
	w/ PDN			3.90	**+7.1%**	97.8	**+6.0%**	77.6	63.9	253.0	**+2.4%**
3D	w/o PDN	540 x 540	57,442	3.02		64.4		78.7	65.7	222.5	
	w/ PDN		47,850	3.64	**+20.5%**	85.0	**+32.0%**	80.1	66.8	245.6	**+10.4%**
14nm node (1.667GHz)											
2D	w/o PDN	375 x 375	-	1.94		71.6		62.2	50.3	192.4	
	w/ PDN			2.09	**+7.7%**	79.6	**+11.2%**	62.5	50.8	201.1	**+4.5%**
3D	w/o PDN	270 x 270	54,000	1.66		57.1		63.4	50.1	178.9	
	w/ PDN		42,543	2.02	**+21.7%**	74.6	**+30.6%**	64.9	51.8	199.7	**+11.6%**
7nm node (2.778GHz)											
2D	w/o PDN	188 x 188	-	0.97		63.2		37.9	29.1	135.7	
	w/ PDN			1.05	**+8.2%**	75.8	**+19.9%**	38.8	29.8	149.9	**+10.5%**
3D	w/o PDN	135 x 135	67,169	0.81		51.9		37.9	29.0	124.3	
	w/ PDN		48,643	1.01	**+24.7%**	70.4	**+35.6%**	38.1	29.3	143.4	**+15.4%**

3.3.1 Impact of PDN

Top metals are usually thick and wide and are not used for signal routing in traditional 2D designs. Therefore, using a significant portion of these metal layers for power delivery does not affect the overall signal interconnect length. The other intermediate metal layers are not heavily used up by PDN (˜20%) and therefore the overall impact, though present, is not extremely high. However, for 3D designs, an important feature is that the top metal of the bottom tier is used for all 3D signal connections (TSVs or MIVs) which connect to the metal1 landing pads of the next tier above. For very few 3D connections, like in TSV-based 3D designs, the presence of blockages in top metal is not very critical. On the other hand, monolithic 3D ICs allow tens of thousands of 3D MIVs and therefore face serious routing blockages when top metal is heavily used up for power delivery. Unavailability of continuous free routing area in these top metal layers results in heavy reduction in MIV count which reduce the 3D benefits.

Figure 16 shows the magnified view of top metal layer of a circuit designed with and without power delivery network. The same regions of the layout are shown in both (a) and (b). The presence of wide power and ground rails prohibits the MIVs from getting freely distributed as explained in Figure 16(c).

3.3.2 PDN impact analysis results

Table 11 and 12 shows the detailed results and comparison of the impact of power delivery networks on 2D and monolithic 3D designs. The total power in any digital IC includes cell-internal power, switching power and leakage power, as categorized during power analysis in Primetime. Switching power can be further divided into cell-pin and wire switching. Cell-pin switching is due to the input gate capacitance of the standard cells while wire switching is from signal interconnects. The individual components are shown in the table to highlight the PDN impact on signal routing power and also the overall power. Wirelength increase is not a direct measure of wire power increase because wire power also depends

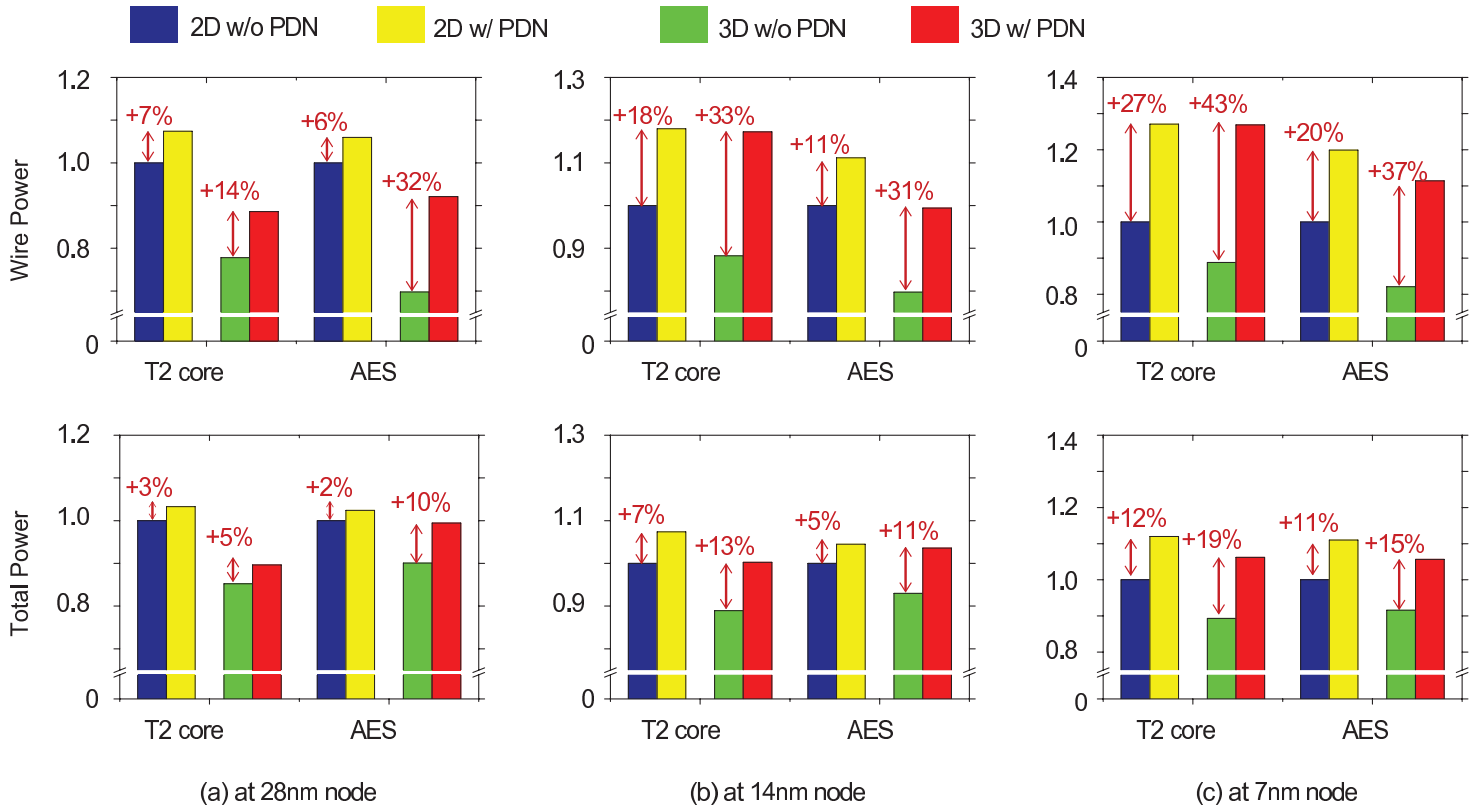


Figure 17: Relative impact of PDN on wire power and total power of 2D and Monolithic 3D designs at (a) 28nm node (b) 14nm node and (c) 7nm node. All values are normalized w.r.t. 2D w/o PDN. Y-axis range is different at different nodes to accommodate the additional impact at advanced nodes.

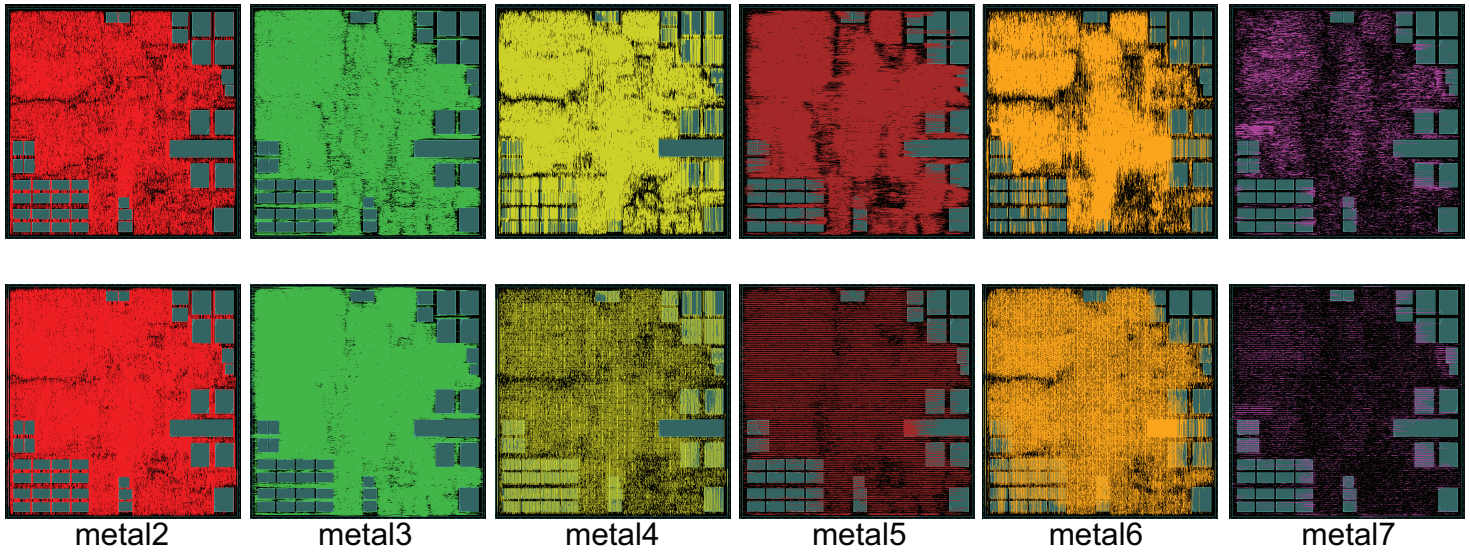


Figure 18: Complete signal routing from metal 2 - metal 7 in the bottom tier (tier with MIVs on top metal) of the two-tier monolithic T2 design. The top row shows the routing done without PDN blockages and the bottom row shows routing after PDN blockages

on the switching activity of the net. Figure 17 highlights the relative increase in wire-power and total-power of both the benchmarks at different technology nodes.

The OpenSPARC T2 single core has memory (register-file) modules and uses up to seven metal layers. Wire Power along with leakage power is a significant portion of the total power. Due to the presence of PDN, the overall increase in power is higher at advanced technology nodes in both 2D and 3D designs. Since interconnect scaling lags behind device scaling, the interconnect impact is more severe at advanced nodes. This impact is much higher in 3D ICs because the MIV locations are affected by presence of PDN in top metal of bottom tier. While the overall power increase due to PDN in 3D T2 is 5% at 28nm, it is as high as 19% at advanced 7nm node. Figure 18 shows the metal usage for signal routing of all the metal layers except metal1 in the bottom tier of monolithic 3D T2 design at 28nm. The bottom tier is critical because its top metal is used heavily for signal MIV insertion (Figure 13). Figure 18 shows that the density of overall routing reduces significantly in metal4 to metal7 due to the presence of PDN. While metal4 and metal7 have the actual power/ground mesh, the presence of PDN via arrays in metal5 and metal6 to connect these PDN wires prohibit continuous signal routing in these layers as well. This can be better visualized by observing the spaced routing in the bottom row of Figure 18 compared to the continuous dense routing in the top row which has routing completed without any PDN blockages. Though, the usage of metal2 and metal3 becomes much more in the designs with PDN blockages, it is not able to compensate for the loss in routing resources in the top metals and also result in increased coupling parasitics. Figure 19 gives a closer look at the metal7 and metal4 layers with and without PDN blockages. The density of routing and the placement of MIVs can be clearly differentiated in the two cases.

AES is a pure logic circuit with no memory modules and uses five metal layers for signal routing. PDN impact on 3D IC power compared to 2D ICs is relatively higher than in T2 because fewer metal layers are used for routing and 3D routing is more affected. The overall power increase due to presence of PDN in 3D IC is 10% at 28nm, 12% at 14nm

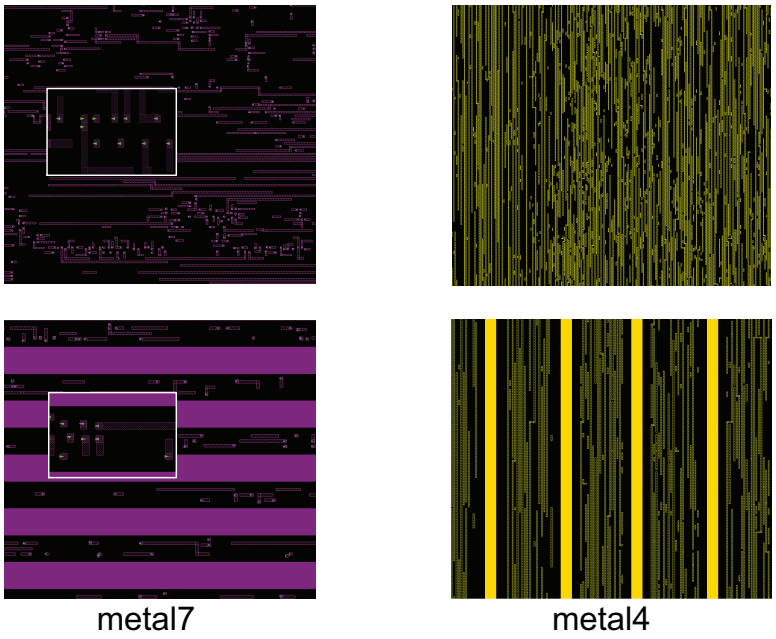


Figure 19: Closer look at the signal routing in the metal layers with the PDN mesh. The top row shows the routing done without PDN blockages and the bottom row is routing with PDN (solid line) blockages

and 15% at 7nm. Interestingly, the relative degradation with advancement of technology is lesser in AES compared to T2. This is because T2 has memory modules and they block a lot of routing area. This additional factor has more severe impact at advanced nodes. AES does not have memory modules and therefore, only interconnect vs device scaling affects the degradation.

3.4 Thermal Impact of Power Delivery Network

3.4.1 New issues

Thermal issues have remained one of the major challenges of 3D IC design. All the potential benefits which 3D IC offers over 2D designs in terms of power savings, wirelength reduction, footprint area reduction and increase in bandwidth are of no use if the chip temperatures are above tolerable limits.

The device layers are extremely thin in monolithic 3D ICs and there is negligible lateral conductivity at the source of power dissipation. Wei *et al.* [54] demonstrated that the

presence of thick PDN metal helps in improving the lateral conductivity and reducing the maximum temperature. However, the power simulations were done at architectural level and do not take into account the fact that PDN present in the individual modules increases the net switching power as demonstrated in the earlier section. This power rise offsets the conductivity enhancement brought about by the thick and wide power/ground wires. This section validates the fact that PDN indeed helps in improving the temperature profile and reducing the maximum temperature, but the overall improvement compared to a no PDN design case is affected by the wire power increase due to increased signal wirelength and coupling.

3.4.2 Temperature results and discussions

Detailed full chip thermal analysis are carried out using ANSYS Fluent and supporting scripts to generate the required input files for Fluent [2]. The thermal analysis tool considers all the layers in the technology and assigns conductivity to each individual tile of the 3D mesh. A full GDS analysis is carried out to know the exact position of the active, poly and signal and PDN metal layers in each tier. For each tile, weighted conductivity is assigned depending on the materials which fall in that tile. For example, a tile may have some portion as $Si0_2$ and some portion as metal and the average conductivity is assigned based on the volume occupied by each material. Conventional heat sink and heat spreader are used on the side of the handle bulk (Fig 13 in flip-chip). The power dissipation occurs only at the device layers of each tier. These power numbers are obtained for each individual cell using PrimeTime analysis.

Three different cases are evaluated for the benchmarks at each technology node. The first case is the 3D design which has no power delivery network. In the second case, 3D design has power delivery network and the corresponding power dissipation numbers, but the conductivity of the metal wires used for PDN is ignored. For the third case, the same design and power as the second case are used, but now the PDN wires are included while

Table 13: Thermal Analysis Results of the 3D designs. Maximum temperature values are reported (in ^oC). Room temperature is 27^oC. % numbers are calculated w.r.t. rise above room in w/o PDN case.

		w/o PDN	w/ PDN	
			w/o PDN conductivity	w/ PDN conductivity
T2 core	28nm	52.81	55.72 (**+11.3%**)	53.99 (**+4.6%**)
	14nm	65.7	71.52 (**+15.0%**)	67.59 (**+4.9%**)
	7nm	80.05	90.21 (**+18.9%**)	83.88 (**+7.2%**)
AES	28nm	58.7	63.89 (**+16.4%**)	62.04 (**+10.5%**)
	14nm	67.61	74.57 (**+17.1%**)	72.29 (**+11.5%**)
	7nm	83.85	95.60 (**+20.7%**)	91.47 (**+13.4%**)

assigning the conductivity values to each tile. This is done by modifications to the post-routing GDS files. The second case is used to isolate the impact of PDN in conductivity.

Table 13 shows the maximum temperature values for each of the evaluated cases at the different technology nodes. The maximum temperature occurs at the device layer away from the heat sink. The full tier temperature maps for these layers at 28nm node are shown in Figure 20. The cooler rectangular regions for the T2 temperature maps are the memory modules in the design which have lower power density compared to the rest of the design. The second and third columns in the figure show that PDN indeed helps in improving the temperature profile by enhancing the lateral conductivity close to the device layers. This is in agreement with [54]. However, the situation is still worse than designs with no PDN in it. This is explained by the fact that there is a power increase from 3D w/o PDN to 3D w/ PDN (Table 11,12). The total power dissipation per unit footprint increases at advanced technology and therefore, the maximum temperature is higher. The relative power increase due to PDN impact is higher for 7nm designs leading to higher increase in temperature even with PDN conductivity. Therefore, while PDN metal does play an important role in enhancement of conductivity, it is very important to include the additional impact of power increase because of PDN.

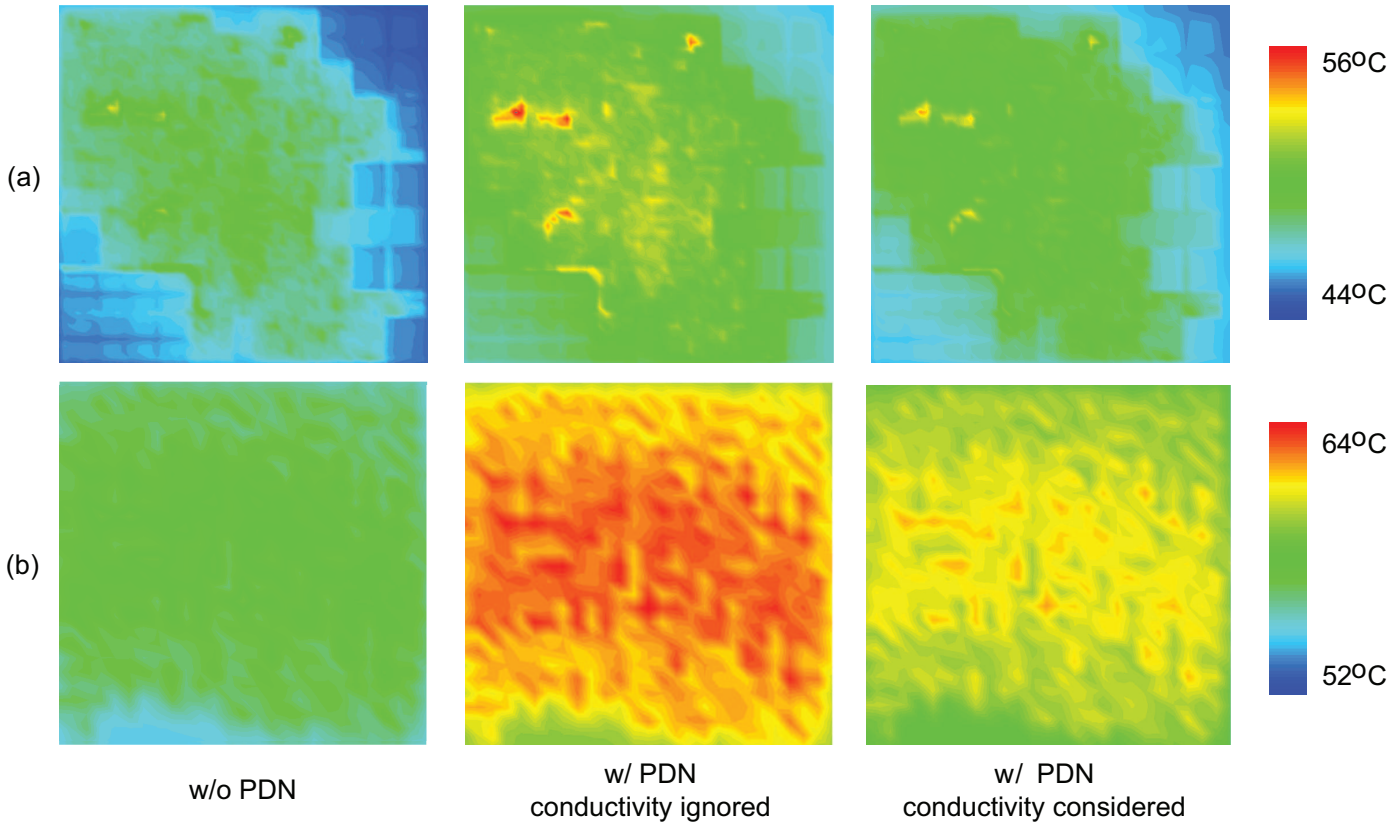


Figure 20: Top Tier (away from heat sink) temperature maps for (a) T2 design and (b) AES design. The actual dimensions are normalized. The middle column is having the same power dissipation as the right column but does not consider the enhancement of conductivity because of PDN.

3.5 Power Delivery Network Optimization

In this section, different power delivery network designs in monolithic 3D ICs are explored to reduce PDN impact on 3D signal routing. The maximum IR drop of the design has to remain within the specified budget while using any other power/ground layout. Simple yet effective changes to the existing PDN are used, which helps in providing more free areas for MIVs in the top metal, while keeping the IR drop under control. The maximum allowed IR drop is set to be around 5% of the supply voltage i.e. 50mV for 28nm PDK, 40mV for 14nm PDK and 35mV for 7nm PDK.

3.5.1 Design styles

Table 14: Summary of metal usage in the various alternative PDN designs (T2 benchmark). All these changes are done to the bottom tier only i.e. the tier with MIV landing pads on top metal.

PDN Design	Metal 7	Metal 6	Metal 5	Metal 4
baseline	40%	-	-	20%
modified	40%	-	-	20%
less topmetal	20%	-	-	20%
multiple metals	10%	10%	10%	20%
intermediate metals	Rings Only	-	40%	20%
no topmetal	Rings Only	-	-	20%

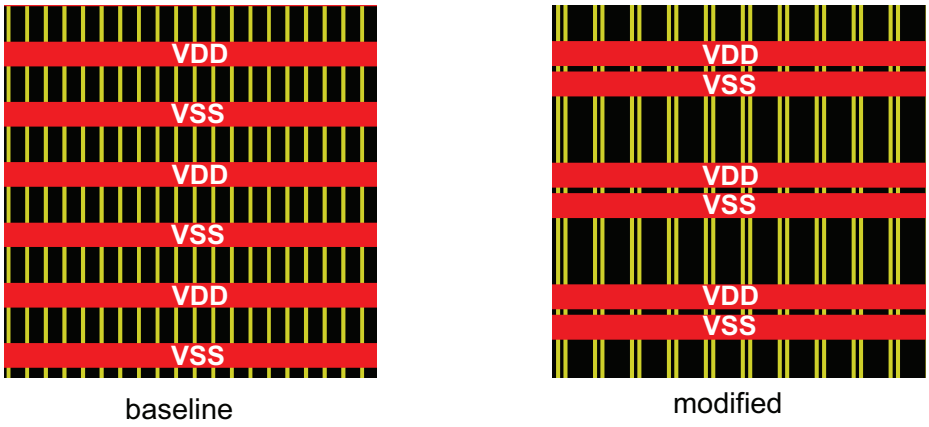


Figure 21: Baseline PDN vs Modified PDN. The extra continuous space between the red top metal wires enhances MIV insertion and routing. The yellow wires are on intermediate metal.

Table 14 summarizes the various PDN design styles used for full chip impact analysis along with the percentage of metal layers used. The first modification is bringing the power and ground rails together instead of having all of them equally spaced. Figure 21 explains this modification. The purpose is to provide more unblocked space for MIV routing and to avoid long detours. This technique is being already used in advanced chip designs but is highly relevant for M3D. The use of this type of technique allows wire power savings of 4.9%, 9.4% and 19.9% in 28nm, 14nm and 7nm designs respectively (Table 15,16). All other design styles discussed later use this modification. This technique can also be used in 2D ICs, but it will not be as effective in bridging the wire power gap because very less top metal is used for signal routing anyways. The other design styles include reduction of top metal layer usage for PDN and compensation with PDN wires in other metals below.

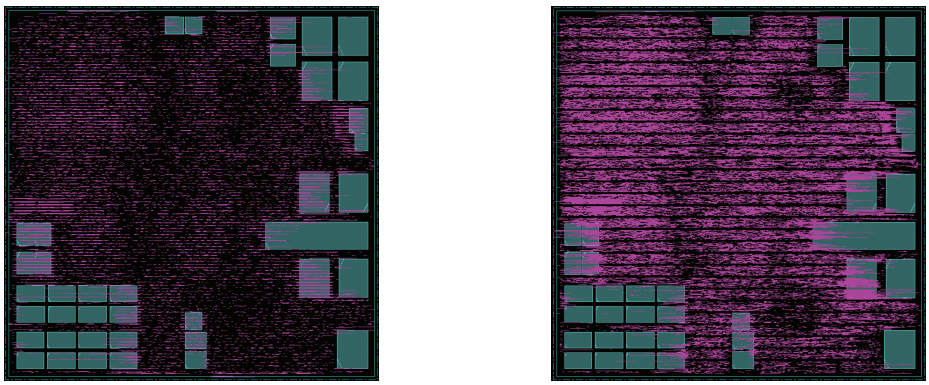


Figure 22: Signal routing for top metal of bottom tier (T2 design) after reducing PDN wires along with clustering of VDD and VSS wires (*less topmetal* design)

The design *less topmetal* reduces top metal usage to half that of the baseline. *Multiple metals* uses the metal layers between the top and intermediate metals used for the PDN in baseline. Only the intermediate metal layers are used in *intermediate metals*. They are connected to the input on top metal through rings only. The final case *no topmetal* uses only one intermediate metal layer for PDN and is connected to supply through rings in the top metal. Even though this is impractical PDN design from an IR drop perspective, nevertheless it is implemented to study the full chip impact for comparison. PDN in metal4

is used as the common intermediate metal used across all design styles. T2 designs use metal7 as top metal while AES uses metal5 as top metal. Therefore, the design styles *multiple metals* and *intermediate metals* do not apply to the AES benchmark as metal4 and metal5 are continuous layers.

3.5.2 Full PDN analysis results

The full chip impact of the different power delivery design styles are summarized in Table 15 and 16. The baseline PDN of the design at the respective technology is treated as reference and all percentage numbers are reported w.r.t this baseline. The impact of PDN on signal routing and power is worse at advanced nodes. Therefore, similar optimization techniques show greater benefits in 7nm designs followed by 14nm and then 28nm designs. The wirelength and total power improvements are more pronounced in the AES benchmark than the T2 benchmark because it is gate-only design. A significant area of T2 is memory modules and related nets are not impacted much by changing PDN layout. Therefore, the improvement in total power is not as high as in AES.

Figure 22 shows the significant improvement in signal routing density in top metal of bottom tier with more continuous space. Not only does it help in adding more vias, it also provides sufficient space for routing without unnecessary detours. The *less topmetal* design shows very good improvement especially in the AES benchmark. In 7nm AES design, 17.6% power is reduced by reducing the density of PDN usage in top metal. For 7nm T2 design case, *intermediate metals* case has maximum power savings of 11.3%. The *no topmetal* design shows best improvement. Though, it is attractive from power perspective, it is not practically feasible due to huge IR drop. All other design styles satisfy the IR drop requirement. The baseline and modified designs have similar IR drop values because the density of the PDN remains the same. Only the VDD and VSS wires come closer to each other. Since there is reduction in power, the maximum temperature reduces as well. Figure 23 shows the AES design temperature maps for the tier away from the heat sink. The

Table 15: Wirelength and Power comparison of various optimized 3D PDN designs for different technology nodes. The footprint area is same for the same benchmark. The % numbers are evaluated w.r.t. the baseline PDN.

PDN Design Style	Wirelength (m)	No. of MIVs	Wire Power (mW)	Cell-Pin Power (mW)	Cell-Internal Power (mW)	Total Power (mW)	IR Drop (mV)	
							Maximum	Average
28nm node								
baseline	17.09	116,727	242.0	158.1	162.7	676.1	29	15
modified	16.86 (**-1.3%**)	118,121 (**+1.2%**)	239.9.0 (**-0.9%**)	158.1	162.7	673.3 (**-0.4%**)	29	15
less topmetal	16.89 (**-1.1%**)	128,715 (**+10.3%**)	237.1 (**-2.0%**)	158.1	162.0	670.3 (**-0.8%**)	55	23
multiplemetals	16.75 (**-2.0%**)	132,205 (**+13.3%**)	236.6 (**-2.2%**)	158.1	161.8	669.9 (**-0.9%**)	39	18
intermediate metals	16.46 (**-3.7%**)	136,156 (**+16.7%**)	237.5 (**-1.9%**)	158.1	161.8	671.0 (**-0.8%**)	32	14
no topmetal	16.08 (**-5.9%**)	137,173 (**+17.5%**)	225.7 (**-6.7%**)	158.1	160.7	657.8 (**-2.7%**)	214	46
14nm node								
baseline	9.91	107,371	161.5	85.8	90.8	406.1	25	14
modified	9.75 (**-1.6%**)	108,048 (**+0.6%**)	149.8 (**-7.2%**)	85.8	88.0	391.6 (**-3.6%**)	25	14
less topmetal	9.77 (**-1.4%**)	111,608 (**+3.9%**)	142.6 (**-11.7%**)	85.8	87.0	383.4 (**-5.6%**)	41	19
multiplemetals	9.69 (**-2.2%**)	114,070 (**+6.2%**)	141.5 (**-12.4%**)	85.8	87.0	382.3 (**-5.9%**)	32	17
intermediate metals	9.65 (**-2.6%**)	114,255 (**+6.4%**)	138.3 (**-14.4%**)	85.8	87.2	379.3 (**-6.6%**)	28	16
no topmetal	9.28 (**-6.4%**)	121,245 (**+12.9%**)	127.3 (**-21.2%**)	85.8	85.4	366.5 (**-9.8%**)	146	44
7nm node								
baseline	5.17	108,068	137.4	49.9	52.8	285.4	22	12
modified	5.02 (**-3.1%**)	108,292 (**+0.2%**)	119.1 (**-13.3%**)	49.9	48.6	262.9 (**-7.9%**)	22	12
less topmetal	5.00 (**-3.3%**)	112,828 (**+4.4%**)	113.3 (**-17.5%**)	49.9	48.0	256.5 (**-10.1%**)	34	20
multiplemetals	4.98 (**-3.7%**)	115,267 (**+13.3%**)	112.2 (**-18.3%**)	49.9	47.9	255.4 (**-10.5%**)	27	16
intermediate metals	4.92 (**-4.8%**)	115,813 (**+7.2%**)	109.8 (**-20.1%**)	49.9	48.0	253.1 (**-11.3%**)	24	14
no topmetal	4.64 (**-10.3%**)	123,081 (**+13.9%**)	100.5 (**-26.8%**)	49.9	47.1	242.9 (**-14.9%**)	118	28

Table 16: Wirelength and power comparison of various optimized 3D PDN designs for different technology nodes. The footprint area is same for the same benchmark. The % numbers are evaluated w.r.t. the baseline PDN.

PDN Design Style	Wirelength (m)	No. of MIVs	Wire Power (mW)	Cell-Pin Power (mW)	Cell-Internal Power (mW)	Total Power (mW)	IR Drop (mV)	
							Maximum	Average
28nm node								
baseline	3.64	47,850	85.0	80.1	66.8	245.6	28	16
modified	3.56 (**-2.2%**)	47,970 (**+0.3%**)	80.8 (**-4.9%**)	80.1	66.6	241.2 (**-1.8%**)	28	16
less topmetal	3.35 (**-8.1%**)	55,311 (**+15.6%**)	73.0 (**-14.1%**)	80.1	66.3	233.1 (**-5.1%**)	45	17
no topmetal	3.35 (**-8.1%**)	56,954 (**+19.0%**)	72.6 (**-14.6%**)	80.1	66.3	232.7 (**-5.3%**)	69	22
14nm node								
baseline	2.02	42,543	74.6	64.9	51.8	199.7	29	17
modified	3.56 (**-2.2%**)	43,813 (**+0.6%**)	67.6 (**-9.4%**)	64.9	49.4	190.3 (**-4.7%**)	29	17
less topmetal	1.81 (**-10.4%**)	51,213 (**+20.3%**)	59.5 (**-20.2%**)	64.9	48.9	181.5 (**-9.1%**)	42	18
no topmetal	1.80 (**-10.9%**)	52,713 (**+23.9%**)	58.8 (**-21.1%**)	64.9	48.9	180.9 (**-9.4%**)	66	23
7nm node								
baseline	1.01	48,643	70.4	38.1	29.3	143.4	25	14
modified	0.95 (**-5.9%**)	49,173 (**+1.1%**)	56.4 (**-19.9%**)	38.1	27.7	127.7 (**-10.9%**)	25	14
less topmetal	0.87 (**-13.9%**)	61,210 (**+25.8%**)	46.8 (**-33.5%**)	38.1	27.8	118.2 (**-17.6%**)	36	16
no topmetal	0.85 (**-15.8%**)	62,691 (**+28.9%**)	46.3 (**-34.2%**)	38.1	27.8	117.7 (**-17.9%**)	53	20

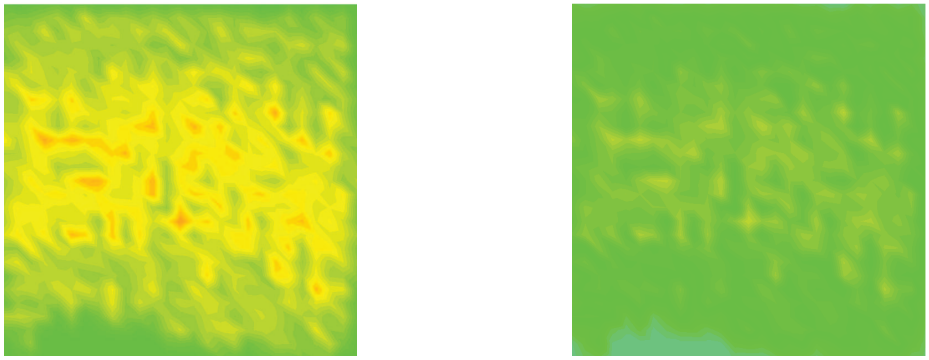


Figure 23: AES top tier (away from heat sink) thermal maps for baseline PDN vs PDN design with less top metal used (at 28nm node). The temperature scale is kept same as Figure 20

left map is the baseline and is same as the one shown earlier (in Figure 20). The thermal map on right in Figure 23 is the temperature map for the *less topmetal* design case.

3.6 PDN Design Guidelines for Monolithic 3D ICs

While designing power delivery network for monolithic 3D ICs, the key is that the top metal of the lower tiers are heavily used for signal MIV landing pads and therefore cannot be fully utilized for PDN as done in regular 2D designs. Therefore, the target is to reduce the usage of top metal for PDN in the lower tiers i.e. tiers grown first in the fabrication process (Fig 13). The proposed design guidelines are as follows:

- The top metal needs to be given more free area for optimal MIV placement and therefore power/ground wires' density in top metal layers should be reduced. The PDN wires removed from top metal can be compensated by using resources in the metal layers below.

- Different power/ground supply wires should clustered together and then these clusters equally spaced on the respective metal layers. This practice is already getting common but needs to be strictly followed in M3D designs. It allows more continuous room for signal routing and MIV planning without affecting IR drop.

- Depending on the footprint and the overall current demand, the pitch of PDN wires can be optimized to reduce the impact of PDN blockages on 3D signal routing while satisfying the IR drop budget.

- PDN needs to be designed much more carefully for interconnected dominated designs compared to cell/memory dominated designs because bad PDN affects signal routing and increases wire power.

- PDN impact on signal power in M3D is worse at advanced nodes and with technology scaling, PDN optimization becomes more critical for M3D designs.

3.7 Summary

This chapter studied the full chip impact of power delivery network in monolithic 3D ICs in detail and demonstrated that the impact is more severe compared to simple 2D designs. The issue becomes much more serious at advanced technology nodes. The role of PDN in the full chip thermal behavior was also analyzed. This study re-validates the fact that PDN does help in enhancing lateral thermal conductivity in monolithic 3D IC and hence results in temperature reduction. But the increase in power dissipation due to increased signal wirelength must be taken into account for accurate temperature analysis, especially in monolithic 3D designs where PDN has more impact. Next, simple yet efficient PDN design styles were evaluated for wirelength and power reduction. This study focuses on designs with single supply voltage. Many practical designs have multiple supply voltages and the top metals are heavily used for PDN layout. With the advent of monolithic 3D IC, it is imperative to find better solutions for such multiple supply designs. This work will serve as a starting reference for further optimization of PDN in such cases, which will help in reducing the impact on signal routing and hence reduce power and maximum temperature.

CHAPTER IV

CAD SOLUTIONS TO HANDLE INTER-TIER VARIATION IN MONOLITHIC 3D ICS

The fabrication process of monolithic 3D ICs lead to inter-tier difference in terms of device performance and interconnect. In the previous chapters on thermal modeling and PDN impact study, it was assumed that all tiers of monolithic 3D ICs have same quality of devices with copper interconnects. The idea there focused on one key aspect while keeping other parameters as nominal. However, for a realistic study and comparison with 2D IC, all designs need to be optimized and evaluated under practical settings, which include the impact of low performance transistors in top-tier and/or the impact of tungsten interconnect in bottom-tier. There has been very little to no CAD research on handling inter-tier performance differences in M3D. This chapter carries out a practical evaluation of M3D benefits under realistic settings and develops CAD methodologies to address these new challenges.

This chapter can be divided into two distinct sections addressing the impact of low thermal budget process and tungsten interconnect, respectively, on the design quality in monolithic 3D ICs (M3D).

In the first section, the impact of tier-to-tier transistor performance difference on full-chip power and performance metrics is quantified. Next, a new Tier-Aware M3D (TA-M3D) design flow is developed that identifies potential timing-critical paths and partitions them into the faster (bottom) tier to minimize the top-tier degradation impact. The unique challenge in timing closure in this case, is how to conduct buffering and sizing on the paths that lie entirely in the top or bottom-tier as well as those that span both tiers. The developed approach handles all three types of paths carefully and closes timing under the given top-tier degradation assumption, while minimizing the total power consumption.

In this second section, tier partitioning strategies are developed to mitigate back-end-of-line (BEOL) interconnect delay degradation and cost issues in monolithic 3D ICs (M3D). First, the routing overhead and delay degradation caused by tungsten BEOL interconnect in the bottom-tier of M3D is studied. Next, two partitioning methods targeted specifically towards BEOL impact reduction are proposed. The path-based approach and the netsize-based approach address the performance and cost challenges while using tungsten interconnect in the bottom tier of M3D.

4.1 Motivation and Background

4.1.1 Low performance transistors in top-tier

Though M3D technology development and design research has gained significant momentum in recent years, there are technological challenges in achieving good transistor performance with low thermal budget [3]. Batude *et al.* have demonstrated low temperature process ($<650^oC$) for transistor fabrication with the measured performance close to that of regular high temperature process transistors ($\tilde{}1050^oC$) [3, 4], but it is more practical to consider some degree of performance reduction in the top-tier devices to have a fair and realistic assessment.

Low thermal budget of sequential device layers mainly affects the dopant activation process, leading to reduced mobility. Prior works have tried Solid Phase Epitaxy at 625^{o}C [4] and laser annealing with low in-depth thermal diffusion [49] for dopant activation. However, there is a performance reduction in the resulting transistors. In terms of design work, Billiont *et al.* proposed a simple design flow using 2D IC tools in [9]. This method folds a 2D placement result along an edge to get 3D designs and does not utilize the true potential of high density MIVs. Chan *et al.* [12] have used Shrunk2D design methodology [45] for their modeling and estimation study. However, the design flow in [45] assumes equal transistor performance in both device layers which can lead to performance failure and optimistic power benefits. Panth *et al.* provide a very detailed analysis of various kinds of

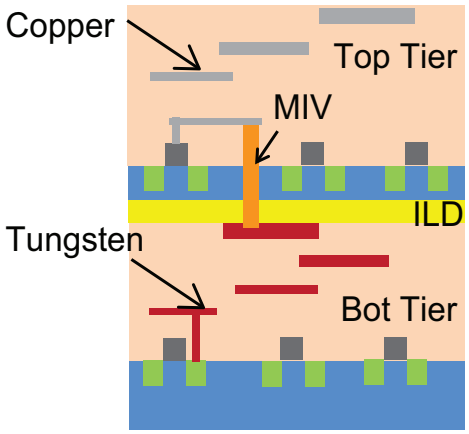


Figure 24: Vertical structure of 2-tier monolithic 3D IC. Tungsten is used for the interconnects in the bottom-tier to withstand high temperature during the top-tier device fabrication.

inter-tier variations for block level floorplanning and design [46]. In this work, the synthesize and layout all possible scenarios for all the blocks is carried out. This incurs a significant design-time overhead and the design approach is conservative. Moreover, gate-level M3D design folding is not addressed.

The major objective is to study the tier-to-tier performance difference in gate-level monolithic 3D ICs and build a robust industry-quality and more practical design flow for the same to achieve maximum power savings with realistic analysis using foundry Process Design Kit (PDK).

4.1.2 Use of tungsten in bottom-tier

Figure 24 shows the vertical layers of two-tier monolithic 3D IC. Because of the thermal restrictions during sequential fabrication, there is the requirement to use tungsten for back-end- of-line (BEOL) in bottom-tier, since copper cannot withstand temperatures close to 650^oC. The bulk resistivity of tungsten (56 Ω.nm) is 3.3X higher than that of copper (17 Ω.nm), and hence has significant impact on performance of the design. Billiont *et al.* developed an EDA methodology to use standard 2D IC tools to obtain M3D designs [9] and studied the impact of tungsten. However, even after optimization, performance is reduced by up to 11% along with increase in power. The methodology proposed in their work just

folds the 2D IC placement results along an edge with very few 3D connections. As a result, the tremendous potential of high density MIVs is not utilized at all and the wirelength and power of M3D designs are higher than even 2D IC designs.

Panth *et al.* proposed the more attractive CAD methodology, the Shrunk2D approach, to obtain significant power reduction in M3D over 2D ICs [45]. Chan *et al.* have used this Shrunk2D design methodology as the "Golden 3D IC implementation" for their modeling and estimation work on monolithic 3D ICs [12]. The Shrunk2D approach allows the use of more MIVs to achieve significant wirelength savings (20-30%) and hence power savings. But it completely ignores the impact of either lower performance devices in the top-tier or tungsten interconnect in the bottom-tier. Both factors cannot be ignored simultaneously for practical M3D technology. Handling the tungsten impact in the design is a major challenge in M3D. While over-designing is the simplest approach, it results in power overhead over 2D ICs [9], therefore nullifying one of the primary benefits of M3D.

Cost is another important factor driving technology progress. The return on investment of using advanced fabrication techniques, multiple patterning etc. is diminishing with technology scaling. To provide a strong contention to an alternative technology or extension to current technology, M3D requires good cost savings along with power savings. Also, M3D fabrication has higher cycle time of fabrication due to multiple layers, even though the footprint is smaller. In addition. robust EDA machinery is necessary to successfully handle the impact of technology and fabrication requirements.

4.2 Low Performance Transistors in the Top-tier

4.2.1 Slow-tier impact study

4.2.1.1 Full-chip design settings

Three cases (-5%, -10% and -15%) of performance degradation in the top-tier over the regular bottom-tier transistors are considered. This covers a broad spectrum including the

Table 17: Cell delay comparison with slower transistors (-5%, -10%, -15%) in the top-tier vs. regular transistors (0%) in the bottom-tier.

Cell type	Degradation			
	0%	-5%	-10%	-15%
INVX1	1.00	1.04	1.08	1.14
NANDX1	1.00	1.04	1.10	1.15
DFFX1 (clk->Q)	1.00	1.08	1.17	1.24

Table 18: Benchmarks used in this low-performance top-tier study.

Benchmark	Frequency	#Cells	#FFs	# I/Os	Type
LDPC	1.87GHz	66K	2,048	4,100	wire-dominated
AES	4.10GHz	166K	10,769	389	gate-dominated
T2 core	1.90GHz	207K	46,732	477	CPU core

advances in low temperature fabrication and helps in assessing the tolerable limits of degradation. Different standard cell timing and power libraries are used after detailed characterization with the respective transistor models having different ON-currents (I_{DSAT}). Table 17 shows the relative cell performance of some basic cells.

This study uses a wire-dominated low-density parity check (LDPC), a gate-dominated advanced encryption standard (AES) and a cpu core (OpenSPARC T2 single core) benchmark to cover the different design categories. Table 18 shows the details of the benchmarks used. All the designs use 14nm finFET PDK at the typical PVT corner and are designed for the same high frequency in all implementations of the given benchmark.

For comparison purpose, the state-of-the-art Shrunk2D design flow [45] is used as baseline. These designs are optimistic because they assume the same transistor performance in both the tiers. Here, the physical dimensions of all cells, interconnects and chip dimension are scaled by $1/\sqrt{2}$ to capture the 50% footprint scaling in the final M3D design. Then, a regular 2D-like design and optimization is followed by localized partitioning into two tiers (to maintain x-y locations of cells), expanding cells back to the original size, monolithic inter-tier via (MIV) planning, and tier-by-tier routing. The individual tier netlists, wire and MIV parasitics, and top-level 3D netlist are used for timing and power analysis. For MIV planning, a 3D metal stack with cell-pins defined in the appropriate metal layers is used,

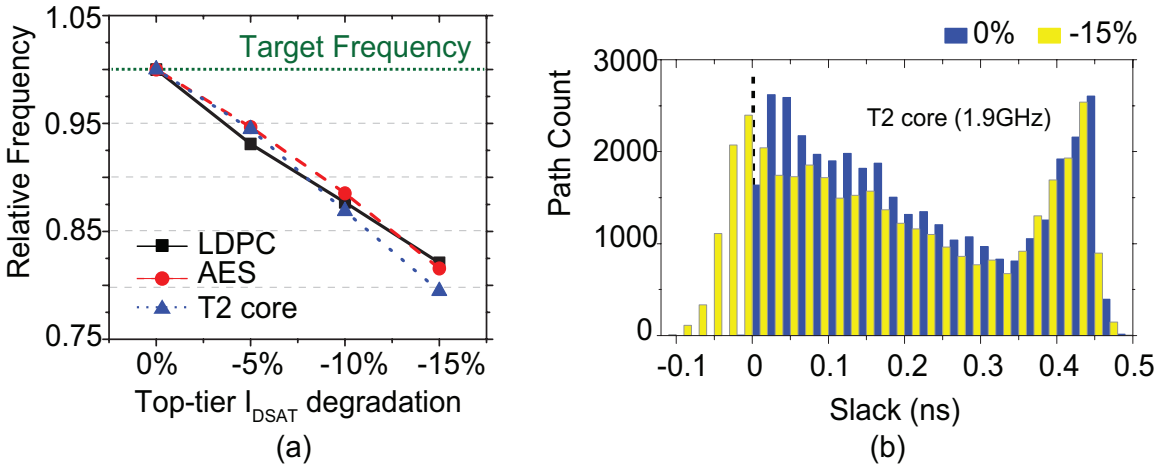


Figure 25: Full-chip impact of slower transistors in the top-tier of monolithic 3D ICs. (a) full-chip frequency degradation, (b) slack distribution of all timing paths in T2 core.

followed by full routing. The locations of the vias going from the top metal of the bottom-tier to the bottom metal (metal1) of the top-tier give the optimized MIV locations [9, 45]. MIV size ($50nm$) and parasitics ($[10\Omega, 0.2fF]$) are determined based on the foundry 14nm PDK via-sizes and the via aspect ratio. The baseline design is 0% degradation (equal-tiers), while the other three cases are named as -5%, -10% and -15% to represent the performance degradation in the top-tier.

For this part of the design work focusing on device performance differences, copper interconnects are used in both tiers. The focus of this work is to study the impact of tier-performance and develop design optimization methods for the same.Power delivery network (PDN) impact study, 3D IC thermal analysis, PVT variation analysis and yield calculations are not included in this section.

4.2.1.2 Full-chip timing impact

The baseline M3D designs are implemented for all the three benchmarks. Baseline (0%) designs, with equal transistors in both tiers, meet the performance requirement. But in practical cases, the same design will experience slower transistors in the top-tier leading to an impact on the overall performance of the design.

Figure 25a shows the relative degradation in performance of the three design benchmarks using the baseline design approach. The system frequency can reduce by 18-21% if the top-tier transistors are 15% slower. Even for 5% slower transistors, the system performance reduces by 6-7%. Figure 25b shows the path timing slack distribution for T2 core design. While the baseline design satisfies timing constraints (blue distribution), the presence of slower transistors in the top-tier leads to a negative slack for many paths (yellow distribution) for the same design, therefore violating the full-chip timing constraints by 21%. For wire-dominated LDPC design, 5%, 10%, and 15% top-tier device degradation leads to 7%, 12%, and 18% full-chip timing violation, respectively. Therefore, designs with prior methodologies, which assume equal quality transistors in both tiers, will fail under practical scenarios. It is imperative to have a robust tier-aware M3D design approach to meet the required performance under practical conditions.

4.2.2 Tier-aware M3D design flow

4.2.2.1 Overview of the new methodology

Timing optimization in commercial tools is carried out in three steps of preCTS, postCTS and postRoute optimization. Majority of path-optimization, buffer addition and cell sizing is carried out during the preCTS optimization stage which includes the parasitics of global routing impact but assumes an ideal clock at all clock-sinks. After full clock tree synthesis (CTS), postCTS timing optimization fixes the resulting clock skew impact. This is followed by detailed routing and parasitic extraction. The residual timing violation cases after full routing are fixed in the postRoute stage.

Figure 26 shows the detailed flow of the proposed Tier-Aware Monolithic 3D IC (TA-M3D) design approach. The idea leverages the fact that most of the timing optimization is achieved in the preCTS stage, and rest of the optimization stages are for fixing any resulting impact of clock/routing etc. In this approach, the optimization tool is provided with information of the new impact of inter-tier performance difference only after the preCTS stage.

Not many timing paths in a design will be violated even with 50% of the cells (top-tier) becoming slower (see Figure 25b). Therefore, it is not necessary to over-design the entire top-tier to meet full-chip performance requirement. Hence, the overhead associated with fixing of the affected paths is minimized by handling them at an intermediate step and not from the beginning.

Scaled physical dimensions of the cells and interconnects are used to capture the impact of reduced footprint and wirelength in evaluating the full-chip parasitic impact, while accommodating all the cells in 50% footprint. Therefore, commercial 2D IC optimization tools can be used to carry out 2D-like M3D optimization. The key to this approach is

- Only regular (bottom-tier) cells are used in the preCTS stage.
- Critical-path based tier partitioning is used to control the impact on timing distribution therefore reducing the additional optimization effort.
- Only the top-tier (slower) cell library is used for CTS and to fix the newly created timing violations after adding the tier information.

Figure 27 shows the layouts for T2 single core 2-tier M3D design.

4.2.2.2 Critical-path based partitioning

The use of only regular (bottom-tier) cells during preCTS optimization stage ensures that none of the paths are over-designed, though some will fail timing after introducing tier information. The output at this stage is the detailed placement and timing slack information with newly added/upsized cells and the design is well-optimized, though not 100%. To minimize the optimization effort and runtime in later stages which include the impact of low performance cells, critical-path based tier partitioning is used right after the preCTS stage. In this scheme, the placement is partitioned such that some of the most critical paths (paths with least positive slack) are intentionally confined in the bottom-tier with minimal change in 2D (x-y) location of all the cells in both the tiers. Though there will still be

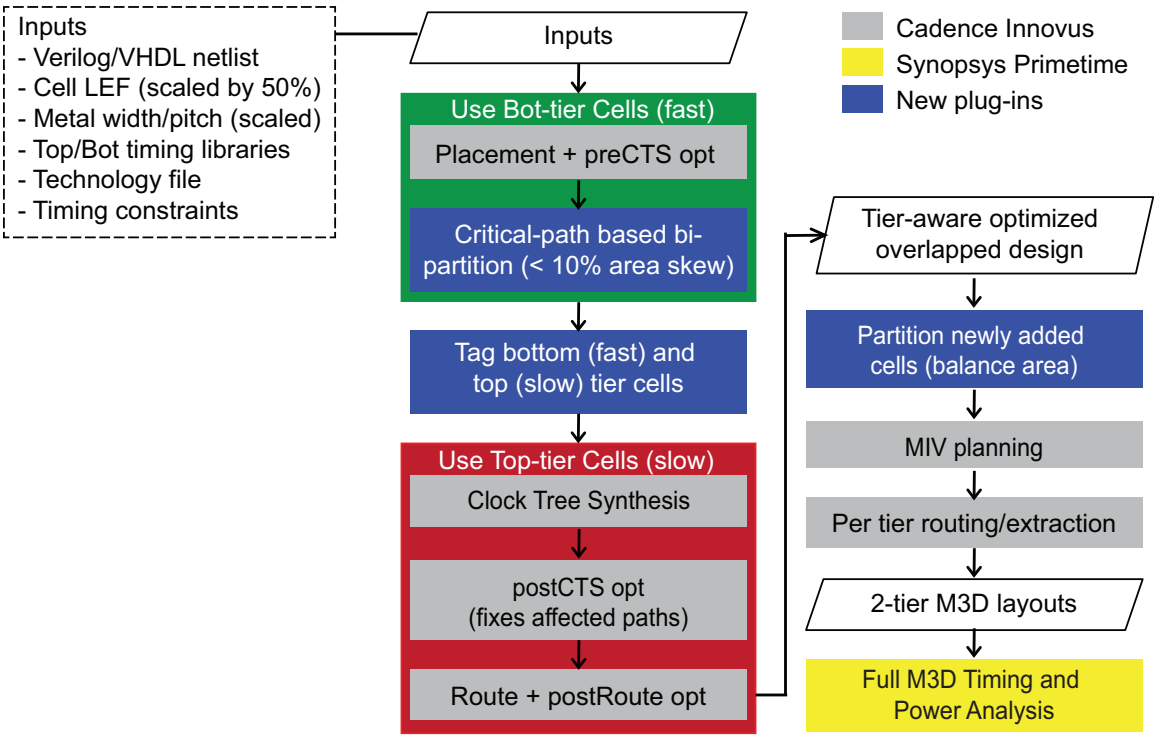


Figure 26: The proposed tier-aware monolithic 3D IC (TA-M3D) design and optimization flow to address slower transistors in the top-tier.

new violated paths appearing due to the top-tier performance impact, it helps in ensuring that the potentially worst paths are avoided from the top-tier. To maintain area balance in both tiers (¡10% area skew), only 10% of the total cells are confined in the bottom-tier based on decreasing order of critical path delays. Localized FM partitioning [25] is used for every $5\mu m$ x $5\mu m$ placement grid in the whole design, with some cells fixed in bottom-tier determined by path criticality. Since MIVs have negligible area overhead, a large total cutsize can be tolerated in this fine grid structure to maintain the optimized x-y placement locations in both the tiers.

The partitioned cells are then marked as per their tier location and this information is provided to the commercial optimization tool which treats them accordingly. Scaled dimensions are continued to accommodate all the cells in half the footprint. The design now represents a placement projection of both the tiers on a single plane but is aware of which cell lies in which tier. The timing optimization tool can then fix the new set of

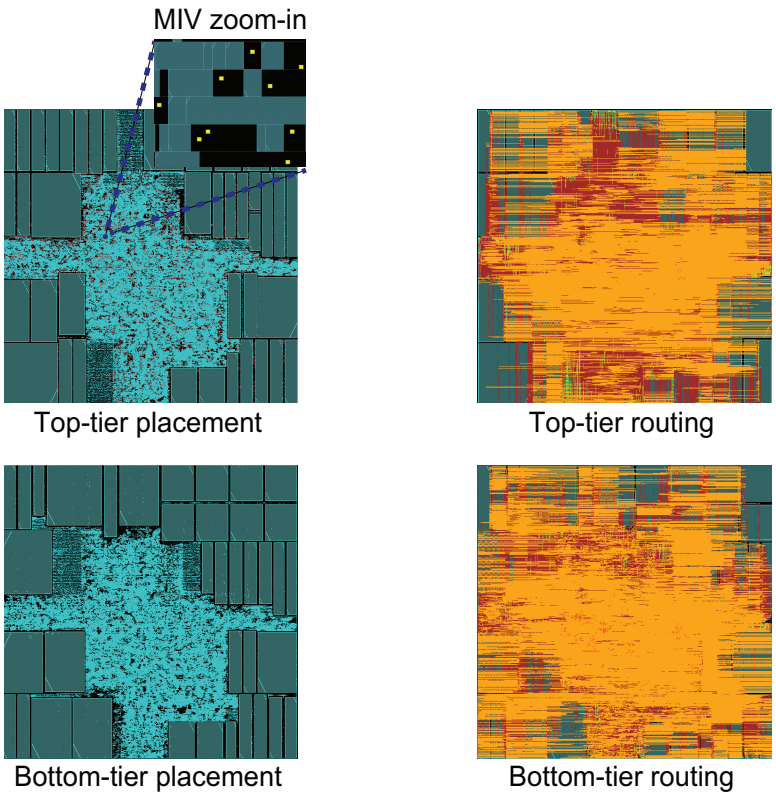
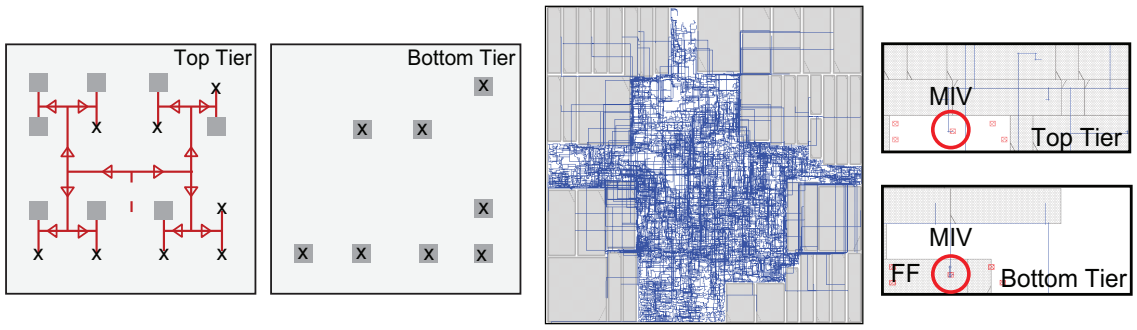


Figure 27: Full-chip monolithic 3D IC layouts of OpenSPARC T2 core using a foundry 14nm finFET PDK. The footprint is 415x415um. Zoom-in shows MIVs (yellow) and cells (cyan).

violating paths only. For subsequent stages after preCTS, only the slower (top-tier) cell library is used. This enables the newly added/upsized cells to satisfy the timing constraints irrespective of which tier they are finally placed on. For most designs, only few paths need to be fixed. and hence, the number of new cell additions is relatively low.

4.2.2.3 M3D clock tree design

Clock tree design is a critical step in any IC design flow and minimizing clock skew across the entire design is an important requirement. Therefore clock tree design is important in M3D as well. Since all the optimization and design is carried out in a 2D-like environment (scaled dimensions) involving cells in both the tiers, the clock tree network is fixed in one tier only to maintain the original optimized tree structure. The MIV connections are planned for clock sinks of flip-flops/register-files in the other tier. Figure 28a shows a



(a) Clock tree synthesis in TA-M3D (x shows clock MIV) (b) T2 Full Clock Tree (top-tier) (c) clock MIV zoom-in

Figure 28: Clock tree synthesis method. (a) FFs in the top-tier are connected with a single tree, and FFs in the bottom are connected only using clock MIVs, (b) Full-chip top-tier clock tree in T2 core, (c) zoom in of a single clock MIV.

simplified version of the clock tree design method.

In M3D designs, I/O access in only through the top-tier which is low performance tier. It is best to fix the clock tree in the tier closest to the I/O i.e. the top-tier. This ensures that there is no additional global clock skew caused by the tree traveling back and forth across the tiers. The final M3D clock tree closely matches the one designed during 2D-like 3D optimization. The clock tree uses only slower cells as clock buffers since it is to be fixed in the top-tier. However, there will be a minor local skew impact at the end points of the tree after partitioning and tier-by-tier routing because the clock-net has to cross the bottom-tier BEOL stack to reach the bottom-tier flip-flop from the clock MIV.

Figure 28b shows the full clock tree for T2 core. While the clock sinks are spread across both the tiers, the primary tree (blue lines) is fixed in the top-tier. There is some clock routing in the bottom-tier due to the clock MIV to flip-flop connections as explained earlier.

4.2.2.4 Rest of the design flow

The inclusion of tier specific cell information affects some paths lying partially or fully in the top-tier. For the bottom-tier cells, the tool only needs to fix clock skew and net routing impact similar to 2D ICs. Since, only the slow (top-tier) cell library is used for later stages, the timing optimization fixes all kinds of paths (confined to one tier or crossing tiers) using

only slower cells. Hence, after a full 2D-like tier-aware 3D optimization, another step of partitioning the newly added cells is carried out. The original faster cells are fixed in the bottom-tier only. The idea is to have area balance and maximize interconnect savings by allowing high 3D cutsize while maintaining the optimized tier-aware 3D placement results. This is followed by scaling up of the cells back to their original size, tier-by-tier placement legalization, MIV planning [9], tier-by-tier routing and power/timing analysis.

4.2.3 Results

4.2.3.1 Power saving challenge with finFET

The total power in a digital design (excluding I/O pads) can be divided into cell-internal, switching and leakage power. Cell-internal power is the power dissipated inside a logic gate (excluding cell pins) due to the switching of internal nodes and short-circuit power. Switching power is further divided into wire switching and cell-pin switching. Wire switching comes from the interconnect capacitance while cell-pin switching is due to the switching of input gate capacitance of the logic gates.

FinFETs have high input gate capacitance and high cell power. For finFETs, cell-power becomes relatively more dominant compared to planar technologies. Table 19 shows the detailed results for all designs implementations for the three different benchmarks. Comparing 2D IC and M3D designs, the key observations are (1) Relative contribution of wire power to total power is lesser for designs using finFET technology. (2) Significant savings in wirelength and wire power in M3D designs. (3) Due to high cell power contribution, total power saving in M3D is modest for AES and T2 core. LDPC is wire-dominated and hence shows significant power savings of up to 28% in M3D. This includes the cell power savings obtained by reducing buffers and cell-sizes.

4.2.3.2 Power vs. performance trade-off

The TA-M3D design flow guarantees timing closed design using high quality commercial tools even with slow transistors in the top-tier, as shown in Figure 29. But to recover the

Table 19: Design comparisons under 0%, 5%, 10%, and 15% top-tier performance degradation. Results with Shrunk2D [45] (which ignores top-tier degradation) are shown for comparison. Leakage power is very small (¡ 1%) at typical PVT corner and hence not reported separately. Power values are reported in mW. Power saving shows the total power saving w.r.t. 2D results.

Design case		Freq. (GHz)	Footprint ($\mu m \times \mu m$)	#Cells (x1000)	Bot:Top Cell Area	#MIVs	WL (m)	Wire Power	Cell-Pin Power	Cell-Internal Power	Total Power	Power Saving
LDPC, wire-dominated												
2D IC		1.87	290×290	65.6	-	-	1.76	127.1	73.8	103.8	305.0	-
Shrunk2D	0%	1.87	205×205	58.0	51:49	22,162	1.22	87.8	54.9	78.0	220.9	**27.6%**
	-5%			62.0	52:48	22,647	1.22	88.1	57.1	81.8	227.1	**25.5%**
TA-M3D Flow	-10%	1.87	205×205	62.5	54:46	22,628	1.23	88.5	59.2	83.5	231.4	**24.1%**
	-15%			70.9	53:47	22,646	1.35	102.1	83.3	103.6	289.1	**5%**
AES, gate-dominated												
2D IC		4.10	320×320	166.4	-	-	1.22	67.4	132.9	181.8	382.7	-
Shrunk2D	0%	4.10	225×225	163.0	51:49	51,051	0.92	59.3	119.5	170.9	350.2	**8.5%**
	-5%			164.3	52:48	51,098	0.96	61.1	129.6	173.2	365.2	**4.6%**
TA-M3D Flow	-10%	4.10	225×225	167.2	52:48	50,921	0.96	62.1	131.5	176.3	369.2	**3.5%**
	-15%			170.6	52:48	51,104	0.97	62.4	136.5	174.8	374.0	**2.3%**
T2 core, processor core												
2D IC		1.90	585×585	206.7	-	-	4.18	144.3	89.9	247.2	483.1	-
Shrunk2D	0%	1.90	415×415	204.7	51:49	68,022	3.32	124.7	88.9	231.4	446.6	**7.6%**
	-5%			204.9	51:49	68,139	3.33	125.1	89.9	229.4	445.8	**7.7%**
TA-M3D Flow	-10%	1.90	415×415	205.0	51:49	68,399	3.33	125.1	90.0	229.3	445.8	**7.7%**
	-15%			205.6	51:49	71,978	3.33	125.8	90.3	229.8	447.3	**7.4%**

lost performance, there is an additional usage of timing buffers and larger cells. While the critical-path based partitioning reduces this usage, too much degradation can lead to a heavy power overhead.

Table 19 summarizes the power-performance trade-off of handling slow transistors in the top-tier of monolithic 3D ICs. Figure 30 shows the effectiveness of the partitioning method in confining most part of the worst paths in the bottom-tier only. The relative presence of worst paths (red lines) in the top-tier of the optimized designs is much lower than that in the baseline case. The cell area ratio is also reported in Table 19 to highlight that area balance is maintained. Note that for similar cell layouts, lower I_{DSAT} implies lesser cell-internal power. However cell-pin power depends on input gate-capacitance of a cell layout and hence remains similar in all cases. Therefore, the overall full-chip power impact is dependent on the total power and path distribution for the baseline (0%) case, and the number of affected paths with a slower top-tier. Figure 31 summarizes the overall power (w.r.t. 2D IC) for the benchmarks under different cases of top-tier performance degradation.

4.2.3.3 Analysis of results

LDPC is a heavily wire-dominated circuit and has only 2,048 timing paths. With -15% degradation in the top-tier, all paths fail to meet timing (Figure 29a) in the baseline design and hence the overhead in fixing them is significant (Table 19). However, for the other cases, the overall power savings is still very high (24-26%) by using the critical path based partitioning. The transition in going from top-tier degradation of 10% to 15% is very sharp due to 14% extra cell usage and the additional wires.

For AES, cell power dominates the total power. Addition of buffers and larger cell sizes used to meet the performance requirement increase this cell power further. Though cell-internal power for cells in top-tier reduce due to slower transistors, cell-count and cell-pin power increase is higher.

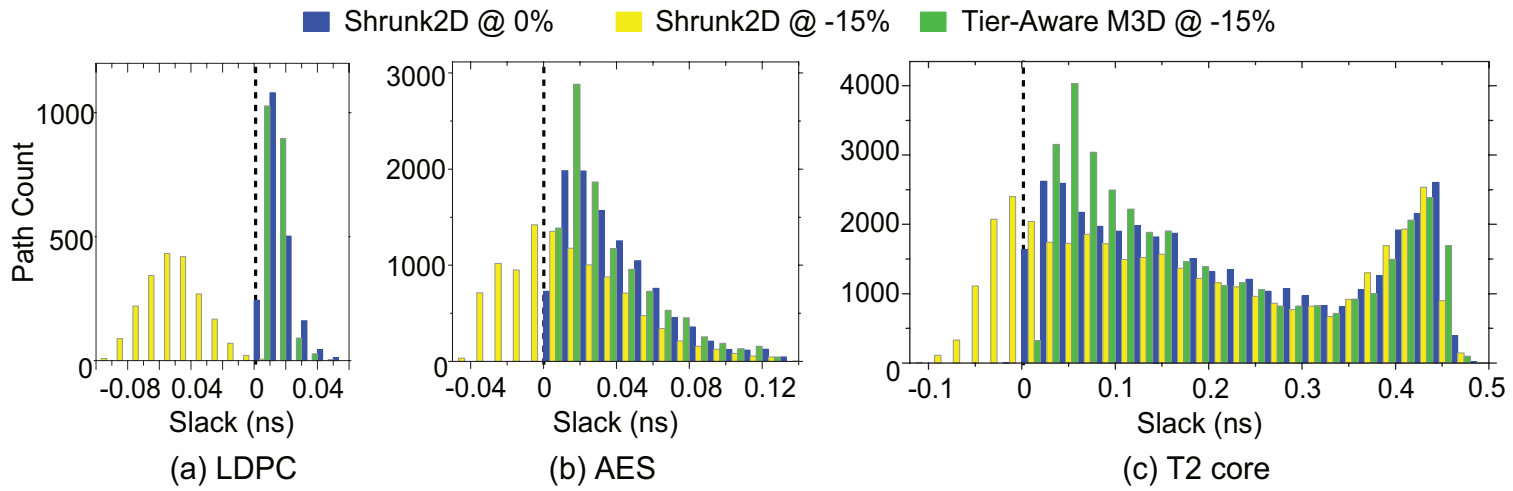
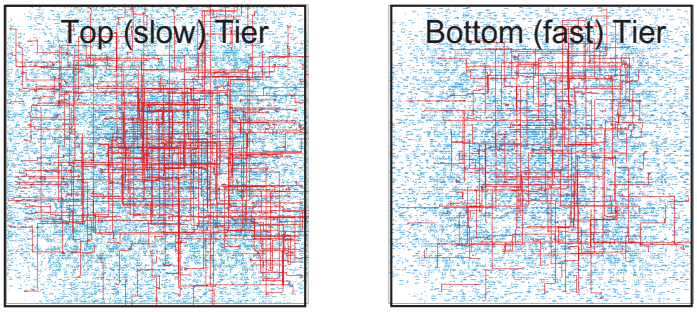
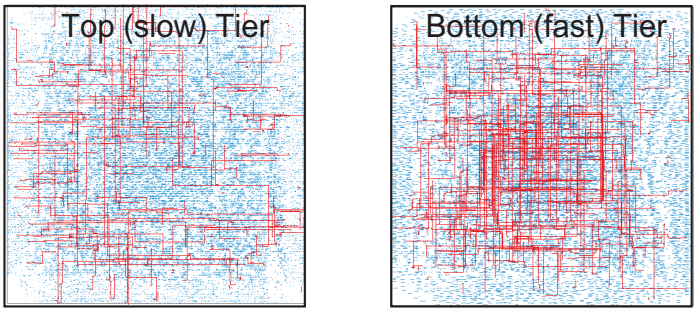


Figure 29: Path delay distributions under 0% and 15% top-tier degradation. All paths to the left of dotted line (= negative slack region) are violating timing constraint. Shrunk2D flow [45] is used for comparison. The paths in design using the new flow satisfy timing even under 15% degradation. (a) LDPC with 2,048 paths, (b) AES with 10,767 paths, (c) T2 core with 38,082 paths.

T2 core has two key design features. (1) It has ~47K flip-flops (Table 18) and ~205K cells. Flip-flops have very high cell power and their count does not change for a given RTL. (2) The timing paths are widely distributed and hence a slower top-tier affects ~15% paths (Figure 29c). Therefore, after fixing the paths using the TA-M3D flow, the relative addition of cells is low, and above that, the top-tier flip-flop power is lower due to slower transistors. Hence the overall power savings remain similar (~7.5%) in all cases.



(a) Shrunk2D Flow [45] (timing violated)



(b) New Tier-Aware Flow (timing closed)

Figure 30: 100 worst timing paths (red lines) in LDPC design under 10% degradation. (a) Shrunk2D Flow [45], timing not closed, (b) The TA-M3D Flow, timing closed. In this design, fewer critical paths are placed in the top (= slow) tier. In addition, excessive buffers and sizing is not done to optimize the slow (= top) tier.

4.2.3.4 Clock tree metrics

Table 20 shows details of the clock buffers and clock power for the different design cases for all benchmarks. LDPC has very few flip-flops (=clock sinks) and hence uses less clock buffers and clock power. T2 core is a cpu core with ~47K flip-flops and 79 register-file modules. Therefore, the clock tree is large (Figure 28b) with many clock buffers. However,

Table 20: Clock buffer count and clock power results. Note that slower transistors in the top-tier show little impact on power due to the small number of buffers added.

	2D	M3D			
		Shrunk2D	-5%	-10%	-15%
LDPC (1.87GHz, 60K cells)					
Clock Buffers	73	77	75	75	79
Clock Power (mW)	2.6	2.4	2.4	2.4	2.4
AES (4.1GHz, 165K cells)					
Clock Buffers	550	356	431	495	504
Clock Power (mW)	25.7	25.4	25.2	25.3	25.9
T2 core (1.9GHz, 205K cells)					
Clock Buffers	1,553	1,475	1,514	1,515	1,515
Clock Power (mW)	66.9	66.0	66.4	66.4	66.3

even with the clock tree using only slower cells, the overall change in power is negligible since the cell-internal power reduction compensates for the extra clock buffer usage. The clock routing required in the bottom-tier is very less compared to the full clock tree and has similar impact for all cases.

4.2.3.5 Runtime analysis

From a design runtime point of view, the only additional design step compared to the prior works [9, 45, 14], is the second partitioning step to determine the tier location of the newly added cells. This additional step has an average runtime of only 63s for LDPC, 463s for AES and 242s for T2 core benchmark. The other runtime increase is the extra time taken by the tool to optimize the newly degraded paths after addition of information about slower (top-tier) cells. However, the overall design cycle for very large designs, even for prior monolithic design implementation flows [9, 45], is few hours (includes 2D-like 3D placement, optimization, routing, partitioning, MIV planning, etc.). Therefore, the addition of a few minutes using the TA-M3D flow is insignificant, especially when the end result is a fully optimized timing-closed design with slower transistors in the top-tier.

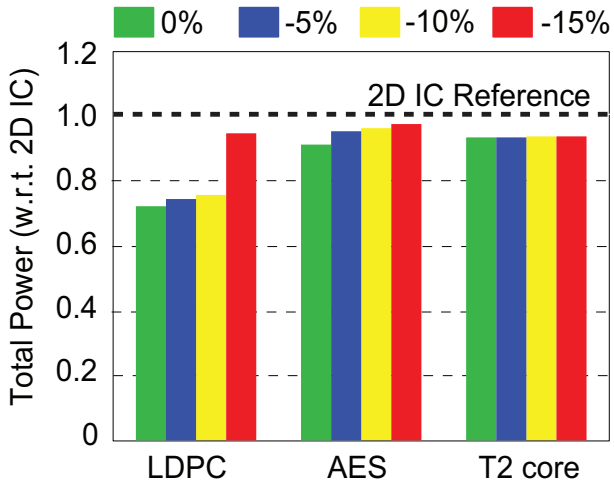


Figure 31: Power consumption (w.r.t 2D ICs) under various top-tier transistor degradation.

4.3 BEOL Impact in Bottom-tier

4.3.1 Full-chip design settings

4.3.1.1 Reference M3D design settings

For the reference M3D design optimization with high quality commercial tools, there are two options, (1) Shrunk2D flow [45] followed by partitioning and (2) Regular 2D IC design with pre-fixed pins, followed by folding along an edge. Shrunk2D flow is shown to have higher power savings [45, 12] and it provides more freedom to implement different partitioning schemes to mitigate technology issues without any over optimization. The second edge-folding technique does not have this flexibility because the partitioning has to be done along edge and cannot be modified. In addition, wirelength and power savings are absent in this approach [9].

Shrunk2D optimization is used for the reference designs. These designs are optimistic because they use same device performance and copper BEOL in both tiers, both of which are not simultaneously practical. In this method, the physical dimensions of cells, wire pitches/widths and chip dimensions are scaled down to create a virtual next node. However, the electrical properties of cells (.lib file) and interconnects (RC per unit length) is kept the same as original technology. These tricks enable the correct mapping of interconnect

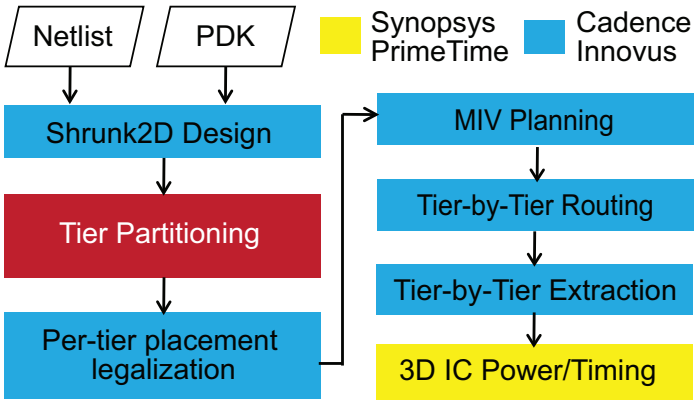


Figure 32: Monolithic 3D IC design flow for impact study of BEOL in bottom-tier. This part of the work focuses on the 3D IC tier partitioning step.

savings obtained by 3D IC designs during 2D-like 3D IC timing optimization. Since RC per unit length (.tch file) is same, the parasitics are correctly evaluated as in a two-tier 3D IC designs but with both the tiers overlapped. The end result is the design with cells in both tiers of monolithic 3D IC projected onto a single 2D plane and well optimized in terms of timing and cell sizing. This is treated as an ideal 3D IC design with zero vertical interconnect overhead.

The next step is division of the 2D placement into a rectangular grid with multiple bins and then partitioning the cells into two tiers using local mincut in each such bin while maintaining decent area balance and similar x-y location as obtained after optimization. MIV planning is carried out by using 3D metal stack of both tiers together with cell pins defined in appropriate locations [9, 45]. The vias running from top-metal of the bottom-tier to bottom-metal of the top-tier give the optimized MIV locations for the provided partitioning solution. Finally, the individual tiers are routed followed by 3D power analysis. MIV size (50nm) and parasitics (10Ω, 0.2fF) are fixed as per foundry 22nm PDK metal pitches, via-sizes, and via aspect ratio. The overall design flow is summarized in Figure 32, with this work focusing on the 3D tier partitioning techniques to mitigate BEOL impact and reduce cost in two-tier monolithic 3D ICs. Equal transistor performance is assumed in both-tiers but with tungsten interconnect in the bottom-tier. Once again, power delivery

Table 21: Benchmarks used in this work. Design metrics are based on 2D IC GDSII layouts using a foundry 22nm FDSOI PDK and commercial CAD tools.

		LDPC	SIMD	AES
		wire dominated	medium	cell dominated
Frequency	(GHz)	1.64	2.36	3.27
Footprint	(μm)	320×320	320×320	300×300
Density	$(\%)$	49.1	70.0	77.6
# Total cells		97,466	150,655	175,548
# Total nets		100,357	154,630	175,808
Metals used	(# layers)	6	6	6
Wirelength	(m)	2.29	2.08	1.19
Wire Power	(mW)	107.4	48.7	31.7
Cell Power	(mW)	117.4	84.8	93.9
Total Power	(mW)	224.8	133.5	125.6
Wire Power %		**48%**	**37%**	**25%**

Table 22: 3D IC design metrics using Shrunk2D flow [45].

		LDPC	SIMD	AES
Frequency	(GHz)	1.64	2.36	3.27
Footrpint	(μm)	198×198	220×220	210×210
Density	$(\%)$	46.2	70.6	76.0
# Total cells		71,355	147,760	174,229
# Total nets		74,779	151,458	174,489

network (PDN) impact study, 3D IC thermal analysis, PVT variation analysis and yield calculations are not included in this section of the research.

4.3.1.2 Benchmarks and metrics

All the designs in this study use foundry 22nm FDSOI PDK at the typical PVT corner and are designed for the same high frequency in both 2D and 3D implementations of the respective benchmarks. Cadence Innovus is used for standard place and route optimization and Synopsys Primetime for power and timing analysis. A wire-dominated low-density parity check (LDPC), a commercial single instruction multiple data (SIMD) engine and a gate-dominated advanced encryption standard (AES) benchmark are used to cover different kinds of designs categories. Table 21 shows the 2D design details of the benchmarks

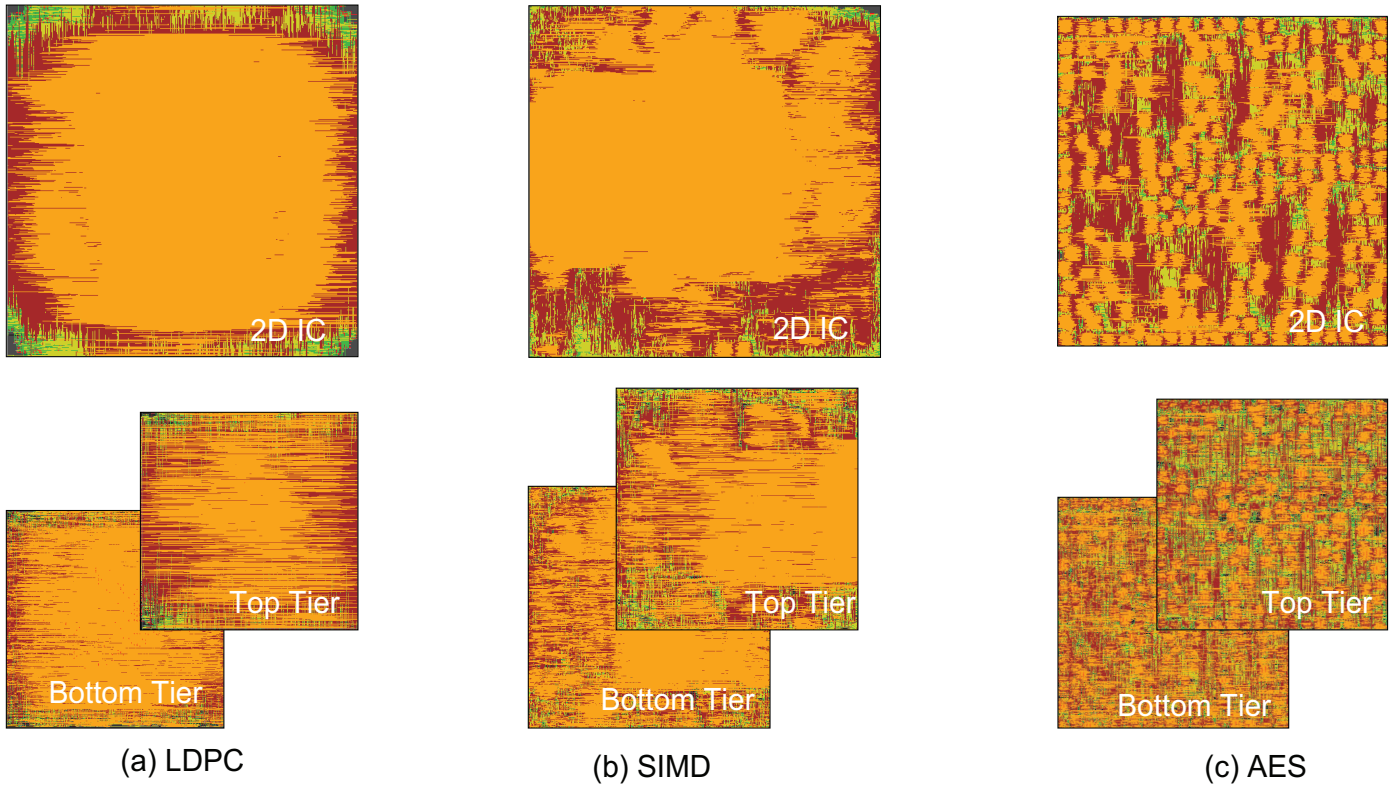


Figure 33: 2D IC and monolithic 3D IC GDSII layouts. Metal6 (topmost metal) is amber color, and Metal5 is maroon. (a) LDPC: very long nets and global spread (b) SIMD: long nets (c) AES: short nets but locally dense. All layouts are to scale.

used in this study. Table 22 shows the high-level design details after 2D-like 3D optimization (Shrunk2D) of the benchmarks. Depending on contribution of wire-power in 2D IC and the wirelength savings in 3D IC, buffering and cell-upsizing is also reduced in 3D IC implementation, leading to cell-power savings as well. The 2D and 3D designs are optimized to have similar final cell-placement density as reported in Table 21 and 22. Figure 33 shows the GDSII level final layouts of the benchmarks. LDPC has very long nets with high wire-power contribution, while AES has shorter nets and localized clusters of dense connections. These factors influence the final power savings observed in the M3D designs, which are discussed in later sections.

4.3.2 M3D bottom-tier BEOL issues

4.3.2.1 Impact of tungsten resistance

Figure 34 shows the delay degradation due to the increase in the resistance of bottom-tier BEOL in the three different benchmarks. The bulk resistance of tungsten is 3.3X higher than copper. The degradation is evaluated at the different points up to 4X worse resistance. Further increase in the resistance will lead to further degradation in timing. In this analysis, the reference design flow is used, while maintaining similar design density in 2D IC and 3D IC designs. For the baseline case, a grid-based regular partitioning on the optimized 3D designs is carried out using Fiduccia Mattheyses (FM) algorithm [25]. During partitioning, the only constraint is to maintain an area-balance with area skew of $<10\%$ across both tiers. Since the 2D-like 3D designs are optimized with copper as interconnects, the impact of tungsten in bottom-tier BEOL is not accounted for during optimization. This leads to significant negative slack in the timing paths which pass through the bottom-tier. The increase in resistance worsens timing in terms of resulting negative slack. However, the relative degradation depends on the original path delay. The degree of degradation also depends on how many timing paths and how much portion of each path crosses the bottom-tier. LDPC is a heavily interconnect-dominated benchmark. The timing paths have longer

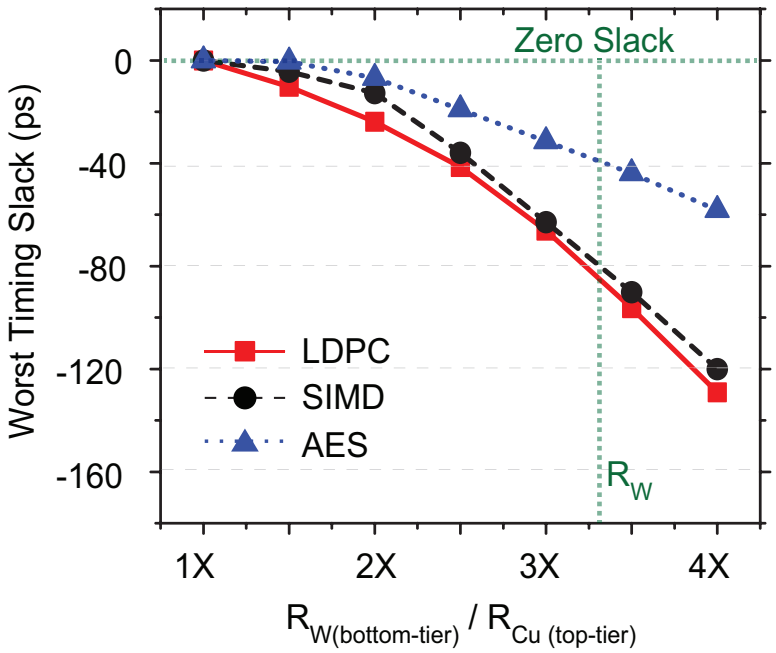


Figure 34: Full-chip timing degradation with respect to increase in the bottom-tier tungsten BEOL resistance. Higher interconnect component in timing paths results in more degradation.

wires compared to AES benchmark. Therefore, the resulting magnitude of negative slack, due to increase in resistance, is highest for LDPC, followed by SIMD and then AES. Simple partitioning schemes have no control on the distribution of these paths. They only consider the connectivity graph and area-balance during partitioning.

4.3.2.2 Accurate routing overhead modeling

Shrunk2D [45] methodology carries out design and optimization in a single 2D plane. However, on splitting the cells into two tiers, the vertical connections between cells have to cross through the entire BEOL stack of bottom-tier before reaching the MIVs in the top-tier. Figure 35 shows the two device layers and the 3D routing in cells across different tiers which has to go through bottom-tier BEOL. With gate to gate 2D distance scaling to sub-micron values in advanced technologies, the 3D routing cannot be considered negligible anymore. This extra routing for 3D nets has two consequences: (1) The timing optimization carried out in the Shrunk2D phase does not account for the additional routing. (2) Extra

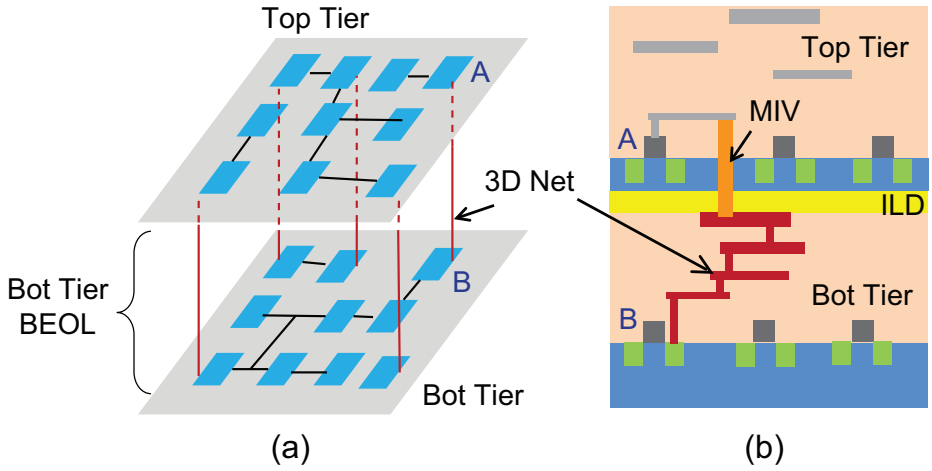


Figure 35: 3D interconnect overhead in monolithic 3D ICs. (a) simplified model of two-tiers with 3D nets, (b) vertical structure showing the 3D routing in bottom-tier. Shrunk2D [45] ignores this overhead

routing means extra interconnect capacitance which results in additional wire power. Both of these issues aggravate further if tungsten is used in bottom-tier. Therefore, reducing metal layers from bottom-tier not only helps in reducing cost, it also helps in reducing negative impact on power and timing. However, directly reducing metal layers without any design consideration will lead to heavy congestion, long detours and a possible routing failure with multiple errors.

The use of tungsten needs to be accounted for during optimization. Also, the 3D routing across the bottom-tier BEOL cannot be completely avoided. However, these adverse impacts can be reduced significantly by using clever partitioning strategy as discussed in the following subsections. The two proposed partitioning methodologies are independently presented without combining one with the other.

4.3.3 Path-based tier partitioning

4.3.3.1 Motivation

The slack distribution of all timing paths of the SIMD benchmark is shown in Figure 36. The slack information is obtained after placement, routing and timing optimization of the 3D designs as discussed in Section 4.3.1.1. The key observation is that the distribution is

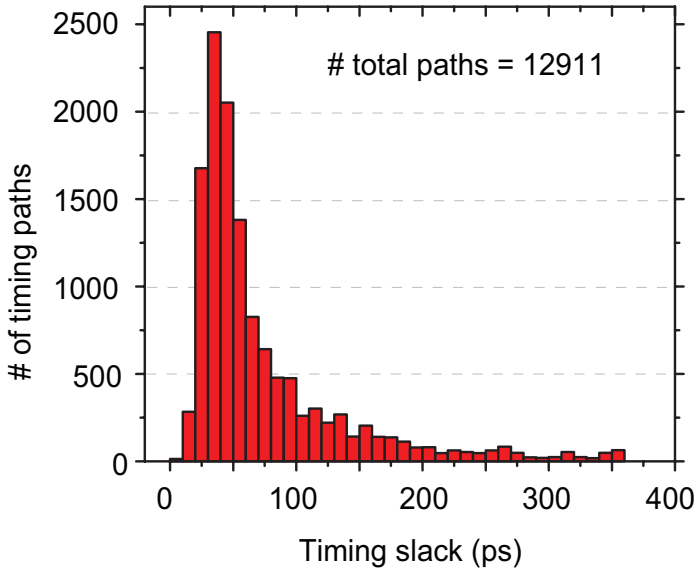


Figure 36: Timing path distribution of optimized 3D design (SIMD benchmark) before partitioning. The wide distribution offers good room of positive slack to tolerate additional interconnect delay.

wide and many paths have very high positive slack. Not all paths are equally critical. In this particular example, almost 50% of the paths have a positive slack of >50ps. Therefore, these paths can tolerate an additional delay of up to 50ps and still satisfy the system timing constraints. The idea uses this fact by confining some of the most critical timing paths in the top-tier of M3D, so that they remain protected from the adverse impact of tungsten interconnect. The tolerance achieved by confining them in the top-tier is influenced by the actual distribution of the timing paths and the degree of connectivity in the netlist.

4.3.3.2 Algorithm and complexity analysis

Algorithm 1 explains the path-based tier partitioning algorithm. The key idea is to try and confine as many worst critical paths in the top-tier as possible without violating area skew constraint and without increasing the cutsize drastically. MIVs are minuscule and occupy negligible area. Therefore, it is possible to increase the cutsize to achieve the target. However, too much increase leads to congestion issues in the bottom-metal (metal1) of the top-tier since all the MIVs need to be routed. The inputs to the new algorithm

Algorithm 1: Path-based tier partitioning algorithm

```
Data: Shrunk2D design (Placement and Timing Slack)
Result: Tier-location of all cells
Function PathBased(path_max)
    GetDesignDatabase();
     path = worst_path;
    while path_count ¡= path_max do
        AssignPathToTier (path, top);
        path_count++;
        path = next_worst_path;
    end
    GridBasedFMPartition();
    if (complete == 0) then
        new_path_max = path_max-20;
        PathBased (new_path_max);
    end
end
Function AssignPathToTier(path, tier_no)
    ∀cell ∈ path : cell →tier = tier_no;
end
Function GridBasedFMPartition()
    if (Bin_Area_skew > 10%) then
        complete = 0;
        return;
    end
    else
        FM partition per bin;
        complete = 1;
        return;
    end
end
```

are already present in the design database after the 3D placement optimization using the Shrunk2D approach. The detailed timing information of all paths is directly obtained from the Cadence Innovus after optimization. Therefore, there is no additional runtime overhead of evaluating timing in this new approach. The *GetDesignDatabase()* function reads the input information into partitioning engine.

The main function *PathBased()* takes in a high max path count as input, and carries out FM partitioning with these critical paths pre-partitioned in the top-tier. Too many paths being fixed in the top-tier may result in high area-skew across tiers. In such a case, the max paths count is reduced by 20 paths and the function is run recursively until the final partitioning result is obtained. The *GridBasedFMPartition()* function divides the layout in a grid structure and carries out mincut FM bi-partitioning ($O(n)$ complexity) in each bin of the grid while maintaining $<10\%$ local area skew across tiers. The major difference from the baseline partitioning (Section 4.3.2.1) is that many cells are pre-partitioned into the top-tier before starting the FM algorithm. The overall runtime depends on the number of recursions occurring to obtain the final results. The recursions can be limited by choosing an aggressive, yet judicious max path count after observing the timing slack distribution.

4.3.3.3 Experimental results

Table 23 summarizes the design results for the different benchmarks using the path-based partitioning algorithm. For comparison, the baseline partitioning results are also presented. One of the very important information provided is the cell-area ratio across tiers. With the new approach, a very good area balance is maintained. High area-skew will result in one-tier requiring more silicon. Due to sequential fabrication process, the other tier has to be of same area, even though major part of it will be whitespace. Therefore, good area-balance is very critical in maintaining footprint reduction benefits in M3D.

Using the developed algorithm, the impact of bottom-tier interconnect resistance can

Table 23: Results with the path-based partitioning. 2.2X to 4X BEOL resistance degradation in the bottom-tier can be tolerated without compromising full-chip M3D power and area balance across tiers. The runtime includes tier partitioning step only.

Method	Bot:Top Cell-Area	#MIVs	Wirelength (m)			Congestion		WireCap (pF)	Power (mW)			**Tolerable Bot-tier R**	Runtime (sec)
			Top-Tier	Bot-Tier	Total	Hor (X)	Ver (Y)		Wire	Cell	Total		
LDPC (1.64 GHz)													
baseline	49:51	19,931	0.72	0.76	1.48	0.01	0.01	244.8	65.4	78.2	143.9	**1X**	45
path-based	53:47	21,803	0.69	0.82	1.51	0.03	0.09	253.8	67.5	78.2	146.0	**2.2X**	66
SIMD (2.36 GHz)													
baseline	50:50	40,558	0.93	0.89	1.82	0.01	0.04	318.6	42.4	80.7	123.1	**1X**	173
path-based	51:49	46,213	0.92	0.95	1.87	0.01	0.08	332.3	44.9	80.7	125.6	**2.9X**	245
AES (3.27 GHz)													
baseline	51:49	35,797	0.54	0.57	1.11	0	0.04	166.5	26.4	89.8	116.2	**1X**	348
path-based	50:50	42,270	0.55	0.65	1.20	0.01	0.11	170.8	27.7	89.8	117.5	**4X**	489

be tolerated to significant extent, without any further timing optimization. The second-last column shows the tolerable limit of worse resistance in the bottom-tier. The baseline designs cannot tolerate any degradation in interconnects. The degree of tolerance depends on the role of interconnect in the benchmarks. LDPC is heavily interconnect dominated and hence interconnect degradation is more critical than in other designs. Even then, up to 2.2X resistance increase in the bottom-tier interconnects can be handled. Further degradation in resistance requires new optimization techniques or new interconnect material innovations. On the other hand, AES has short nets and hence interconnect impact is relatively lower. With this developed approach, full 4X resistance increase in the bottom-tier interconnect can be handled.

The side-affects of achieving the desired goals is the increase in cutsize (hence MIV count) and congestion which leads to more wirelength with minor increase ($<2\%$) in total power. The congestion in both horizontal and vertical direction is also shown in the Table 23. The runtime for path-based tier-partitioning is higher because the partitioning process goes through a few recursions, depending on choice of initial max path count to be fixed on top-tier. However, the overall runtime for partitioning is a few minutes only and is negligible compared to the total design runtime of few hours (includes MIV planning, tier-by-tier routing and parasitic extraction). Therefore, path-based partitioning algorithm proves highly beneficial to reduce or completely remove the optimization overhead of handling tungsten interconnects.

4.3.4 Net-based tier partitioning

4.3.4.1 Motivation

While CAD methodology and power reduction for monolithic 3D ICs has been extensively studied [9, 45, 12], there are no prior works targeted towards saving cost in M3D to push it further as an alternative to technology scaling or extension of current technology node. Nayak *et al.* [41] have modeled power-performance-cost (PPC) benefits of M3D and

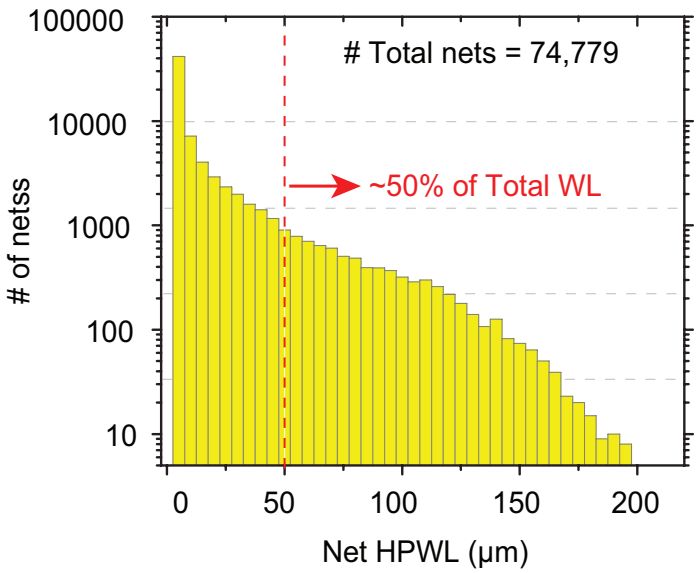


Figure 37: HPWL distribution of all nets in optimized LDPC M3D design before partitioning. Longer nets add up to 50% of total HPWL, although their count is lower than shorter nets (y-axis is in log scale).

TSV-based 3D ICs, but only with estimated cost benefits and no actual designs. In this subsection, the focus is on reducing the cost in M3D by proposing a net-based partitioning algorithm with the objective of reducing metal layer usage in the bottom-tier. M3D gives significant savings in wirelength. Therefore, for the same design, the usage of metal layers in one or both the tiers can be reduced. This helps in reducing overall fabrication cost of an IC, by reducing the number of masks and cycle time. The reduction is significant especially in advanced nodes where tight minimum pitch and multiple patterning increase the back end of line (BEOL) cost by a considerable amount.

4.3.4.2 Algorithm and complexity analysis

Algorithm 2 describes the developed net-based tier partitioning algorithm for gate-level M3D. The target is to reduce the metal usage in bottom-tier as much as possible without affecting the cell-area balance between the two tiers and while minimizing power overhead. Initial overlapped 2D placement results provides a clear idea of the x-y locations of the cells but not the tier of placement. Figure 37 shows the HPWL distribution of all nets in

the LDPC benchmark after 3D design, but before any partitioning. The longer nets, though relatively lesser in count, account for 50% of total wirelength.

The idea is to choose the longest nets from these 2D placement results and force such nets in top-tier only. The other shorter nets and the resulting 3D nets use the bottom-tier metal with reduced demand of routing resources. Note that by placing the long nets in top-tier, any extra wires are not added, as the placement locations of cells are already determined. In fact, it actually helps in reducing the 3D routing overhead by avoiding unnecessary snaking of wires across two tiers for longer nets, which may be cut during simple area-balance partitioning only. The routing resource demand in the bottom-tier is also reduced as relatively shorter nets are routed in the bottom-tier.

There are two good ways to determine the length of a net. They are net's cell-count i.e. number of cells in the net and the 2D half perimeter wirelength (HPWL) which is the Manhattan distance between the placement location of the extreme cells in the net. A net with many cells may not be necessarily spread across a larger area with longer wirelength, depending on how the design was optimized and the cells placed. Therefore, it is not the best metric to assess the length of a net. Therefore, the HPWL data obtained from the Shrunk2D placement results is used. The partitioning tries to maintain the 2D locations of the cells. Therefore, HPWL gives a practical estimate of the net wirelength excluding the 3D routing overhead. The calculation and sorting of HPWL of all nets is a one-time requirement just after Shrunk2D placement. Since the cell 2D placement information is already available, time complexity is $O(nlogn)$, where n is the total number of nets.

The main function *NetBased()* picks all nets with their HPWL greater than a threshold HPWL value ($hpwl\_max$), and fixes the cells in these nets in top-tier. The parameter $hpwlmax$ is recursively reduced until an area-balanced partitioning solution is achieved. The final cell and net count in each tier is similar to that in normal area-balanced partitioning but the longer nets get confined to the top-tier. The runtime overhead is negligible compared to the runtime of entire process of monolithic 3D IC design (Figure 32) which

Algorithm 2: Net-based tier partitioning algorithm

Data: Shrunk2D design (Placement and Nets' HPWL)
Result: Tier-location of all cells

```
1  Function NetBased(hpwl_max)
2      GetDesignDatabase();
3      ∀n ∈ Nets such that n →hpwl ¿= hpwl_max    AssignNetToTier (n, top);
4      GridBasedFMPartition();
5      if (complete == 0) then
6          new_hpwl_max = hpwl_max-5;
7          NetBased (new_hpwl_max);
8      end
9  end
10 Function AssignNetToTier(net, tier_no)
11     ∀cell ∈ net : cell →tier = tier_no;
12 end
13 Function GridBasedFMPartition()
14     if (Bin_Area_skew > 10%) then
15         complete = 0;
16         return;
17     end
18     else
19         FM partition per bin;
20         complete = 1;
21         return;
22     end
23 end
```

involves much more time intensive steps of design optimization and routing.

4.3.4.3 Experimental results

Using the net-based partitioning algorithm, monolithic 3D IC metal layer savings are achieved for the three different benchmarks. Table 24 shows the detailed results in terms of per-tier wirelength, congestion, wire capacitance, wire-power and total power with different metal layer usage for different design implementations. Cost savings by reduction of metal layers cannot be disclosed due to foundry confidentiality requirements. One key advantage of the net-based algorithm is that, the wirelength per-tier can be significantly skewed without affecting the area-skew. This is because the long nets are picked and fixed on the top-tier leading to controlled wirelength skew across tiers. As a result, the routing demand in the bottom-tier is reduced significantly, resulting in up to three metal layers reduction. The number of metal layers used across the various implementations are shown in the second column of Table 24. All the 2D IC and baseline M3D implementations of the three different benchmarks use 6 metal layers (both tiers of M3D). The cell-area ratio in top and bottom-tiers is also shown to highlight the fact that there is a good area-balance across the tiers to maintain footprint savings. The cutsize increases with higher wirelength skew across tiers.

As a consequence of reducing routing resources, the overall relative demand of routing resources increase, which results in more congestion and higher total wire capacitance due to increased proximity of signal wires. Simply reducing the metal layer limit in the bottom-tier for the same baseline partitioning, (i.e. without any wirelength skew) makes the bottom-tier unroutable. This is because, the routing demand in bottom-tier remains the same as the baseline case but with much lesser resources. However, this approach intentionally creates the wirelength skew, while maintaining area-balance. Therefore, the metal layers and hence cost is reduced with minor power overhead. Also, the routing in the top-tier metals becomes denser resulting in more wire capacitance. Depending on the design characteristics, there are varying results in terms of reducing number of metal layers

Table 24: Results with the net-based partitioning and metal layer saving in the bottom-tier. Top-tier uses six metal layers in all cases. 3 metal layers are reduced in all cases with minimal impact on full-chip M3D power and area balance across tiers. The runtime includes tier partitioning step only.

Method	# Metals (Bot-Tier)	Bot:Top Cell-Area	#MIVs	Wirelength (m)			Congestion		WireCap (pF)	Power (mW)			Runtime (sec)
				Top-Tier	Bot-Tier	Total	Hor (X)	Ver (Y)		Wire	Cell	Total	
LDPC (1.64 GHz)													
baseline	**6**	49:51	19,931	0.72	0.76	1.48	0.01	0.01	244.8	65.4	78.5	143.9	45
net-based	**5**	51:49	20,110	0.87	0.61	1.48	0.01	0.01	250.3	66.7	78.5	145.3	45
	4	52:48	21,871	0.97	0.52	1.49	0.01	0.06	265.3	71.2	78.6	149.8	51
	3	55:45	23,301	1.09	0.43	1.52	0.14	0.07	278.3	74.8	78.6	153.4	81
SIMD (2.36 GHz)													
baseline	**6**	50:50	40,558	0.93	0.89	1.82	0.01	0.04	318.6	42.4	80.7	123.1	173
net-based	**5**	50:50	45,386	0.99	0.78	1.78	0.06	0.07	320.8	42.8	80.7	123.5	210
	4	52:48	52,178	1.13	0.68	1.81	0.07	0.19	343.7	46.1	80.8	126.9	225
	3	54:46	50,274	1.28	0.66	1.94	0.29	0.26	385.5	52.2	81.1	133.3	275
AES (3.27 GHz)													
baseline	**6**	51:49	36,069	0.54	0.57	1.11	0	0.04	166.5	26.4	89.8	116.2	348
net-based	**5**	49:51	42,270	0.56	0.56	1.12	0.01	0.06	170.2	27.2	89.8	117.0	442
	4	55:45	38,993	0.69	0.53	1.22	0.17	0.18	184.2	32.6	90.3	122.9	482
	3	55:45	44,893	0.82	0.47	1.29	0.27	0.28	198.6	34.8	90.5	125.3	528

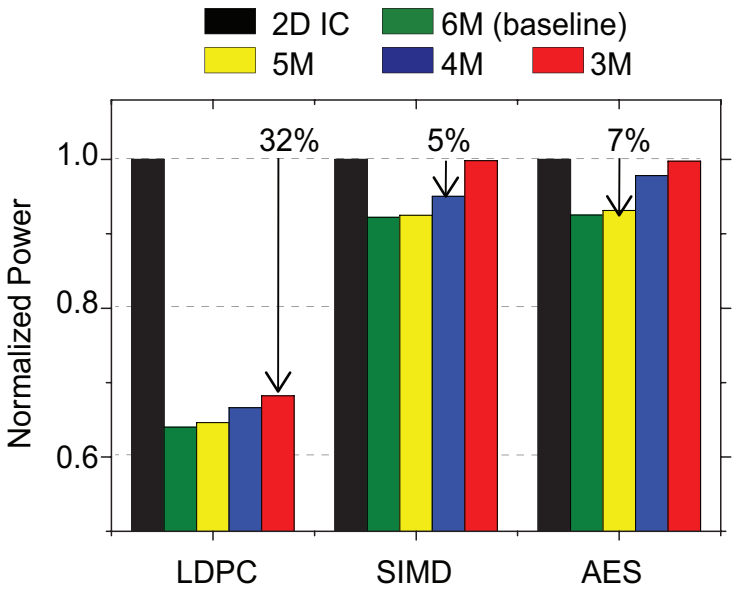


Figure 38: Normalized power comparison of 2D IC, baseline 3D IC and net-based partitioned 3D IC with reduced metal layers in the bottom-tier. Top-tier has six metal layers in all cases.

in the bottom-tier vs. power increase. Using only two metal layers in the bottom-tier leads to very heavy congestion and incomplete routing. Moreover, power delivery requires the use of some intermediate metal layers. Hence, the partitioning methodology is evaluated to usage of three metal layers in the bottom-tier. The normalized power saving comparison is shown in Figure 38. All values are normalized to the 2D IC power of the respective benchmark.

LDPC is an interconnect dominated benchmark with long nets having global spread. With this approach, three metal layers are reduced in the bottom-tier. The routed layouts of both tiers are shown in Figure 39. The congestion increases from 1% in the baseline case to 14% in this case. However, the power benefits over 2D IC is still a significant 32% compared to the 36% in the baseline case. While wire savings reduce with more congestion, cell-power savings remain almost constant across all implementations, resulting in relatively lower impact on total power savings. For the SIMD and AES, benchmarks, the nets are relatively shorter and localized as was shown in Figure 33. Cell power savings are comparatively much lower in these designs. Therefore, any impact on wire-power savings

reflects heavily on total power savings. The localized congestion spots results in relatively higher power increase for the design implementation with three metal layers in the bottom-tier. The overall power savings are modest 5-9% for the different cases of bottom-tier metal layers reduction. The partitioning runtime becomes higher to obtain larger wirelength skew because the starting $hpwl\_max$ is lower and multiple recursions occur before obtaining an area-balanced partitioning result. However, as discussed earlier, this increase is negligible compared to the total design runtime of few hours which includes MIV planning, tier-by-tier routing and parasitic extraction.

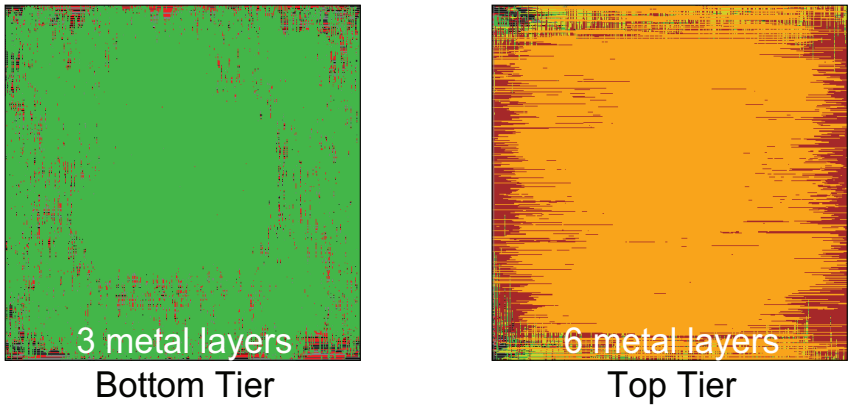


Figure 39: M3D layouts for LDPC benchmark with three metal layers in bottom-tier, using the net-based partitioning.

In general, usage of more metal layers helps in increasing the power savings due to the relaxed routing conditions. This net-based tier partitioning algorithm helps in keeping power in check, while reducing the number of metal layers in the bottom-tier. With the net-aware tier partitioning methodology, wirelength skew with proper area balance is achieved. Therefore, the congestion in bottom-tier is reduced significantly and this helps in error-free routing with low wire capacitance. Though the power-only savings are higher while using more metal layers, the combined savings including cost of reduced masks is higher with reduced metal layers in bottom-tier.

4.4 Summary

The full-chip impact of slower transistors in top-tier and tungsten interconnect in bottom-tier were independently studied. Overlooking the performance degradation during design process can result in full-chip timing failure. A new critical-path based Tier-Aware M3D (TA-M3D) design flow was developed to handle such slow transistors in the top-tier using industry-quality tools. The critical-path based partitioning and design approach ensures minimal design time and power overhead. Up to 26% power savings in M3D was demonstrated for wire-dominated benchmarks with slower transistors in the top-tier. While the design flow is robust and ensures timing closure even with a 15% degradation of the top-tier, the study also shows that for the used PDK, up to 10% degradation in the top-tier is well tolerable to maintain similar power savings as a no-variation design case.

Next, the critical issues of BEOL impact on the performance of gate-level monolithic 3D ICs was addressed. A path-based tier partitioning algorithm was developed to handle the impact of increased resistance of the bottom-tier interconnects with negligible design overhead. A tolerance of up to 4X resistance increase in the bottom-tier interconnect was demonstrated, without any additional timing optimization steps. This was followed by the development of a strategy to reduce the cost of monolithic 3D ICs by reducing the number of metal layers in the bottom-tier. The net-based tier partitioning algorithm helps in reducing the number of metal layers in bottom-tier without any routing congestion, therefore enabling cost reduction. Using this algorithm for two-tier monolithic 3D IC for an interconnect dominated benchmark, up top three metal layers in the bottom-tier were reduced, while saving 32% power compared to 2D IC.

CHAPTER V

MONOLITHIC 3D IC TECHNOLOGY POWER-PERFORMANCE-COST COMPARISON

In general, 3D ICs have been shown to provide attractive solutions to extend Moores Law [55, 4, 46]. In 3D ICs, vertical vias can vary in size from 5μm (TSVs) to 0.05μm (MIVs), offering a wide range of granularity in vertical connections. TSVs are used in block-level connections, whereas MIVs have a potential to offer vertical connections with a density reaching over 10 million/mm2. With increased challenges in scaling device technology to below 10nm nodes, M3D is being carefully studied and evaluated as a feasible alternative to scaling or an extension to an existing technology node. But any future generation technology node requires reduction in power, savings in cost, and improvement in performance. Therefore, as a feasible and attractive alternative to scaling of devices, M3D needs to provide overall benefits in terms of power and cost reduction. In addition, the magnitude of power savings of M3D is heavily dependent on the selection of device technology and process design kit (PDK) as well.

In this chapter, a comprehensive study of power, performance, area, and cost comparisons among TSVs, mini-TSVs (TSV with smaller diameter), and MIVs is presented. A design comparison of 2D ICs, monolithic 3D ICs and TSV-based 3D ICs is carried out using a silicon-validated foundry technology and commercial quality designs. Through full-chip layouts and sign-off analysis using commercial-grade tools, the potential of monolithic 3D IC is explored and validated in terms of power, performance, area and cost (PPC) against that of TSV-based 3D ICs and 2D ICs. Next, the impact of transistor technology on the power savings in monolithic 3D ICs over traditional 2D ICs is discussed.

5.1 Monolithic 3D IC vs TSV-based 3D IC

TSVs are few micrometers in diameter and they have large pitch (30-50μm) and keep-out-zone (KOZ) requirements. In addition to that, they have large parasitic capacitance. With logic gate size scaling to less than 0.5μm^2 in 14nm technology nodes and below, such large TSVs will be beneficial only for coarse-partitioning for block-level or die-level memory-on-logic 3D IC designs. This limitation prohibits the design of TSV-based 3D ICs to maximize interconnect and cost savings with fine-grained partitioning for 3D ICs. Monolithic Inter-Tier Vias (MIVs) are similar to metal-to-metal vias in dimensions and parasitics, enabling very fine-grained 3D partitioning in both the gate level as well as intra-gate level i.e. transistor level [37]. Miniscule 3D vias are necessary in advanced technology nodes for fine-grained partitioning.

This section compares the power and area benefits of TSV-based 3D ICs and monolithic 3D ICs w.r.t. 2D ICs for the same OpenSPARC T2 benchmark design. The study is based on full RTL-GDSII layouts using a silicon-validated foundry 14nm FinFET PDK.

5.1.1 3D IC technology scaling impact

TSV-based 3D IC technology increases memory bandwidth, improves system performance, and enables heterogeneous integration. However, as an alternative approach to scaling below 10nm technology nodes, TSVs are too big. While advanced node (14nm and below) standard cells are below 0.5μm^2 in area, the via diameter in most advanced TSV technology is still few micrometers with every TSV surrounded by a Keep-Out-Zone (KOZ), where no transistors can be placed. They also have minimum pitch requirement of few micrometers to tens of micrometers depending on TSV size and technology. Therefore, for advanced technologies, the area overhead associated with TSVs is too high for fine grained logic on logic partitioning.

Figure 40 shows the form factor comparison of 3D vias with logic gates of 14nm and 28nm NAND gates. The TSVs are actually aggressively sized to 2μm for mini-TSV and

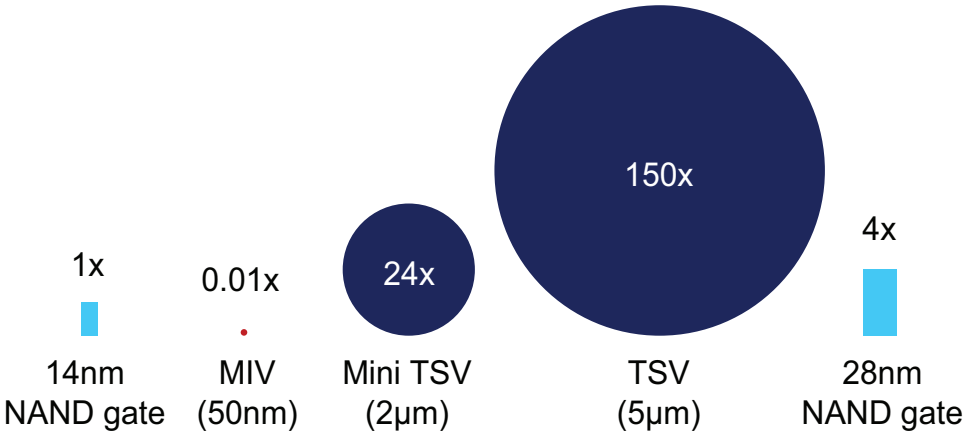


Figure 40: Relative size comparison of 3D vias and NAND gates (14nm and 28nm). The diameter of monolithic inter-tier via (MIV) is $50nm$, mini TSV is $2\mu m$, and TSV is $5\mu m$.

$5\mu m$ for TSV. A practical TSV diameter would be 5-10μm. Also the figure does not show the KOZ which is a huge additional area overhead. Though TSV sizes may be reasonable for older technologies, they are too big for fine-grained partitioning of 3D IC designs in 14nm FinFET technology. However, MIVs are of similar sizes as regular vias and are perfectly suitable to enable gate-level as well as transistor-level partitioning in 14nm technology. The following sub-sections show the direct comparison of the area and power impact of both kind of 3D IC technologies.

5.1.2 Design methodology and setup

In this comparison study, OpenSPARC T2 single core is implemented in 2D IC, TSV-based 3D IC and monolithic 3D IC. For simple power and area comparison, all three design implementations are targeted for 1.5ns clock period (667MHz frequency) at typical operating conditions. TSV-based logic-on-logic 3D ICs cannot operate at very high frequencies due to large TSV parasitic impact. Figure 41 shows the layouts of the different implementations with a section of TSV and MIV placement enlarged. The GDSII layouts for 2D IC were designed using Cadence Innovus. while timing and power analysis is carried out using Synopsys PrimeTime. For 3D IC designs, the following specific design methodologies were used.

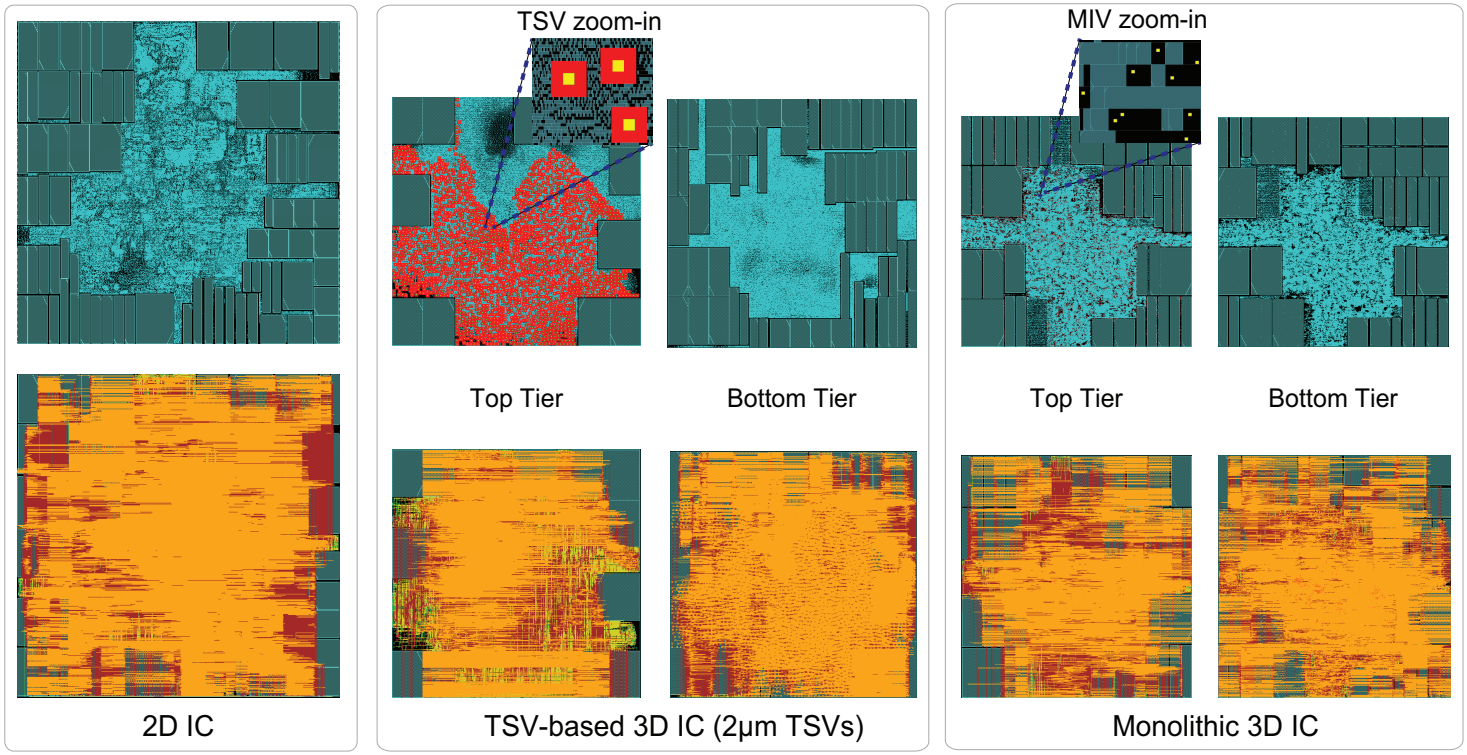


Figure 41: Commercial quality GDSII layouts of OpenSPARC T2 single core using a foundry 14nm FinFET PDK. The footprints of 2D, mini TSV 3D, and monolithic 3D IC (M3D) are $585 \times 585\mu m$, $450 \times 450\mu m$, and $415 \times 415\mu m$, respectively. The red region around yellow TSV is the Keep-Out-Zone (KOZ). Note that we use a much deeper zoom-in in M3D to reveal MIVs, so cells shown in cyan colored rectangles appear larger than in TSV zoom-in.

5.1.2.1 TSV 3D IC design

TSVs are much larger compared to the 14nm logic gates. Therefore for practical and optimistic analysis, the TSVs are aggressively sized to $5\mu m$ and $2\mu m$ diameters and pitch of $15\mu m$ and $6\mu m$, respectively. The extent of KOZ around a TSV is kept equal to its diameter. The methodology developed in [34] is used for TSV-based designs. The partitioning tool *hmetis* [1] is used to partition the gates in the synthesized netlist into two dies using global mincut. To accommodate the large size of TSVs, area-unbalanced partitioning with global mincut is used. This minimizes the TSV count which are planned in the die which has lesser number of cells. The degree of unbalance is more for $5\mu m$ TSV design because the area overhead associated with TSVs is much larger. After partitioning, the TSVs are treated as large cells in top-tier and detailed 3D IC cell placement gives us optimized TSV locations [34]. The pitch and KOZ requirements for TSVs are honored by placement blockages around TSV pads (Figure 41). This is followed by die-by-die legalization of cell placement and TSV location and then detailed routing in Innovus. Since the timing constraints for TSV interface across dies are not known initially, a few iterations of die-by-die place and route followed by timing analysis and budgeting are required to successfully close timing for the entire 3D IC design. For 3D IC timing and power analysis, a TSV RC of $[0.01\Omega, 20fF]$ and $[0.05\Omega, 3fF]$ for $5\mu m$ and $2\mu m$ TSVs respectively is used.

5.1.2.2 M3D design

Shrunk2D design flow in [45] is used for the monolithic 3D IC designs. In this technique, physical dimensions of all standard cells, interconnects and layout sizes are scaled by $1/\sqrt{2}$ to represent 50% footprint scaling but all electrical properties are kept the same. This is followed by regular placement, timing optimization and routing using commercial 2D IC tool. The end result is the design with cells in both tiers of monolithic 3D IC projected onto a single 2D plane and well optimized in terms of timing and cell sizing. This is treated as an ideal 3D IC design with zero vertical interconnect overhead. The next step is division

of the 2D placement into different rectangular bins and then partitioning of cells into two tiers using local mincut in each such bin without heavy area imbalance or change in 2D placement location. The idea is to maintain the optimized design placement but legalize the physical dimensions by dividing the cells into different tiers. Partitioning is followed by MIV planning by using a 3D metal stackup with cell pins defined in appropriate metal layers [45] and then routing of the 3D design. The locations of vias going from top metal of bottom tier to bottom metal (metal1) of top tier give the optimized MIV locations. After tier-by-tier routing and interconnect parasitic extraction, the individual tier netlists, wire parasitics, MIV parasitics ($[10\Omega, 0.2fF]$) and top level 3D netlist is used in Primetime for timing and power analysis.

5.1.3 Full-Chip design comparison

Table 25 shows the comparison of TSV 3D ICs and monolithic 3D IC in terms of area, wirelength, 3D interconnect overhead and power savings w.r.t. 2D IC implementation. Note that placement density for designs are similar with TSV designs having slightly higher utilization. Hence, the comparison of 3D via overhead is fair in terms of silicon area required for full design. Here placement density represents the portion of total silicon area used for logic gates, memory modules and 3D vias with their KOZ. Even optimistic TSV sizes result in area overhead for 14nm T2 core. With 996 and 1,839 TSVs, there is a 76% and 18% area overhead for $5\mu m$ and $2\mu m$ TSV 3D IC designs, respectively. Table 25 also shows the area overhead of the 3D vias and their KOZ. The 3D via and cell size comparison is magnified in Figure 41.

Due to less number of TSVs allowed (high area overhead), the wirelength reduction is not significant in TSV-designs compared to that in M3D and hence power savings are very low. Monolithic 3D IC implementation has 21% wire savings and 9% power savings over 2D ICs, most of which come from wire-power savings. The wire power savings are dependent on both wire capacitance and switching activity of the respective nets. The

Table 25: Area, 3D interconnect overhead, and power comparison of the designs shown in Figure 41. The frequency of operation is $667MHz$. The numbers is parenthesis are relative to 2D IC values. Placement density in 3D ICs is average between two tiers and includes area used by 3D vias.

		2D IC	TSV 3D IC ($5\mu m$)	TSV 3D IC ($2\mu m$)	Monolithic 3D IC
Footprint	(μm)	586×586	550×550 (-12%)	450×450 (-41%)	415 x 415 (-50%)
Silicon Area	(mm^2)	0.344	0.605 **(+76%)**	0.405 **(+18%)**	0.344 **(0%)**
Cell-utilized Area	(mm^2)	0.260	0.257 (-1%)	0.260 (0%)	0.258 (0%)
Total Wirelength	(m)	4.20	4.86 (+16%)	4.30 (+2%)	3.30 **(-21%)**
# 3D Vias		-	996	1,839	48,790
3D Via Pitch	(μm)	-	15	6	0.1
KOZ (around 3D Via)	(μm)	-	5	2	0
Area Overhead	(mm^2)	-	**0.241**	**0.069**	**0.001**
3D Via overhead %		-	**39.8%**	**17.0%**	**0.3%**
Placement Density		75.6%	82.1%	81.2%	75.8%
Wire Power	(mW)	**54.3**	62.5 (+15%)	53.1 (-2%)	43.4 **(-20%)**
Cell Pin Power	(mW)	**30.7**	30.4 (-1%)	31.3 (+2%)	29.1 **(-5%)**
Cell Internal Power	(mW)	**77.9**	77.0 (-1%)	80.8 (+4%)	75.4 **(-3%)**
Leakage Power	(mW)	**1.22**	1.15 (-6%)	1.22 (0%)	1.15 **(-6%)**
Total Power	(mW)	**164.1**	171.0 **(+4%)**	166.4 **(+1%)**	149.1 **(-9%)**

Table 26: Estimated metal layer usage in three different 3D Via count options for a 1 million gate design.

Via Type	many (~100K)	medium (~10K)	few (~1K)
Bottom-tier metal layers			
M3D	2	2	2
mini-TSV	full-1	full	full
TSV	full+1	full	full
Top-tier metal layers			
M3D	full-2	full-1	full
mini-TSV	full-1	full	full
TSV	full+1	full	full

leakage is very low in all designs because power is analyzed at room temperature under nominal conditions.

5.1.4 PPC analysis

In addition to the size of inter-tier vias which results in area overhead, the PPC quality of 3D ICs also depends heavily on the total via usage that is determined by tier-partitioning. In each 3D IC via option shown in Table 26, three different via usages are used: few, medium and many refer to the total via count relative to the total gate count in the design. For monolithic 3D, cost advantage is clearly seen, as the bottom tier uses only 2 to 4 metal layers. The other options (TSV and mini-TSV) use almost all metal layers (full). For the many TSV option, number of metal layers needed can be high (full+1), because the footprint becomes excessively large.

Figure 42a shows PPC comparison for M3D using three via count options. Assuming we achieve a 40% power reduction with M3D (many and medium) against its 2D IC counterpart, PPC gain becomes equivalent to one technology node advancement. That means an M3D design built in 14nm technology node with highly optimized die partitioning and physical design can match the PPC performance of a 2D IC built in 10nm. Going from medium to many has little impact on M3D PPC. However, when the via count reduces from medium to few, PPC drops by 10%.

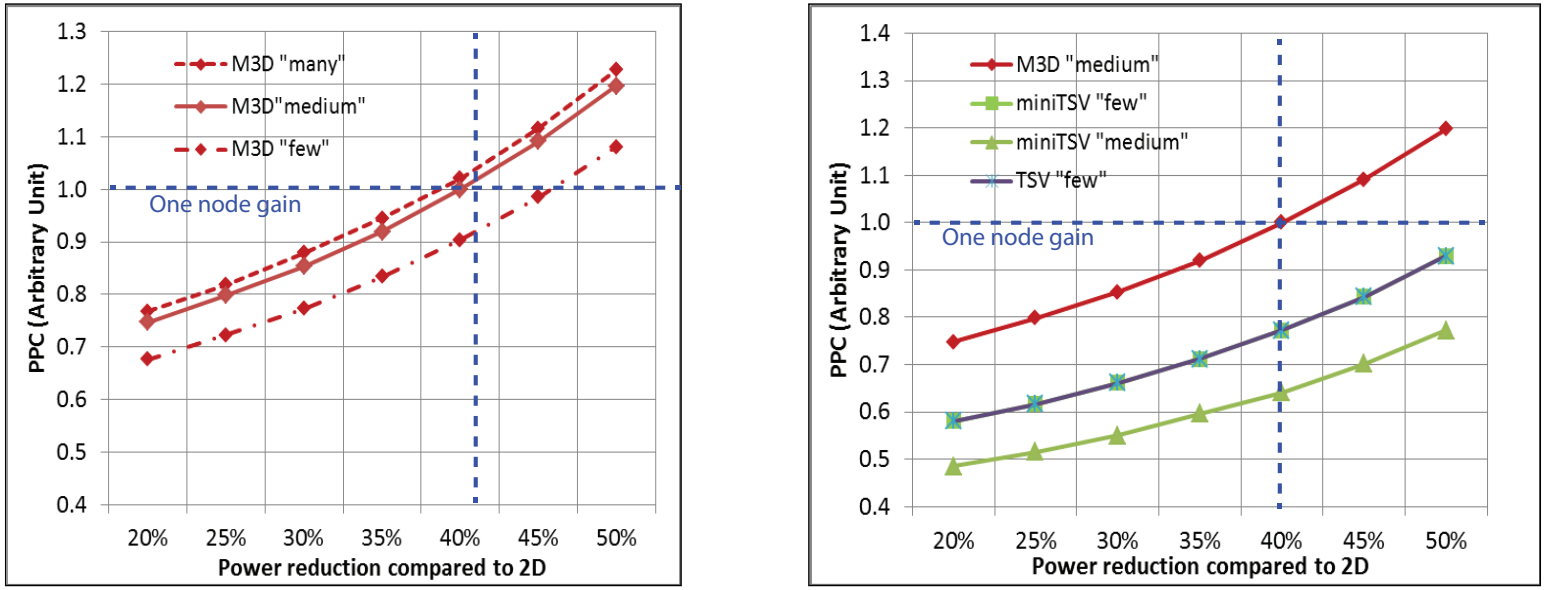


Figure 42: PPC comparison (a) for monolithic IC under three via counts. (b) among M3D, mini-TSV, and TSV.

Three different via technologies in Figure 42b. Mini-TSV and TSV delivers only a half-node PPC advantage even with rigorous power optimization. The PPC sensitivity to via count in mini-TSV is opposite to that found in M3D case (Figure 42a). PPC value in mini-TSV degrades by 17% when more vias are used, which is mainly due to its area overhead. Both TSV and mini-TSV show comparable PPC if their via usage is few. This is because both design options do not exploit the full benefits of 3D IC vias.

5.2 M3D Across Device Technologies

In this section, the impact of transistor technology on the power savings in monolithic 3D ICs over traditional 2D ICs is compared. The results are based on gate-level 3D IC partitioning and full RTL to GDSII design and analysis of a Low Density Parity Check (LDPC) benchmark circuit block with use of two different silicon validated foundry technologies. These two technologies have the same nominal operating voltage, but differ in terms of device performance, power, and gate capacitance.

5.2.1 Background

Power in any digital integrated circuit can be divided into switching power, cell-internal power, and leakage power. Switching power can be further divided into switching of wires and switching of cell pins i.e. input gate capacitance of cells. Cell-internal power is the power dissipated inside of cells due to switching of internal node capacitance (excluding cell pins) and short circuit power during operation. Therefore, total power comprises of wire-switching power, cell-pin switching power, cell-internal power, and leakage power. M3D implementation helps in significantly reducing wire length and hence wire-power due to footprint reduction compared to 2D ICs. In addition, there is cell savings in terms of lesser timing buffer usage and use of smaller cell sizes because of reduction in back-end loading. The weighted sum of savings in these different power components contribute to the total M3D power savings.

Table 27: Normalized M3D power comparison across two different device technologies

Power Component	Technology 1		Technology 2	
	2D IC	M3D IC	2D IC	M3D IC
Wire switching	0.36	0.26	0.43	0.26
Cell-pin switching	0.27	0.21	0.25	0.17
Cell-internal	0.37	0.29	0.31	0.23
Total	1.00	0.76	1.00	0.67

5.2.2 Technology details

Two different silicon validated Foundry PDKs are used in this design study with LDPC. Both the technologies have a nominal operating voltage of 0.8V. They are referred to as Technology 1 and Technology 2, respectively. Technology 1 is used as the baseline during normalized comparison. Figure 43a-b show the power consumption and stage delay, respectively, of a minimum size inverter chain with fan out of 3. At the nominal voltage, Technology 2 consumes just 0.3x power of Technology 1 but is 2.7x times slower. Therefore, Technology 1 has higher drive strength but very high cell-internal power. The slopes of the curves in Figure 43b are also different indicating that the threshold voltage of Technology 2 is higher and therefore delay increases sharply with reduction in supply voltage. Figure 43c compares the input pin capacitance of three different standard cells of different drivability in the two technologies. Technology 2 has 35% lower pin capacitance compared to Technology 1. The contribution of interconnect vs. cell power in a digital integrated circuit is heavily dependent on these factors. Overall, Technology 1 has higher performance, higher power, and higher input pin capacitance, while Technology 2 is a low power technology, with relatively lower input pin capacitance.

5.2.3 Results

Both 2D and M3D designs are designed for iso-frequency operation in the respective technologies and the relative savings in M3D are compared. Table 27 summarizes power in LDPC benchmark designed with these two different technologies. The power numbers are

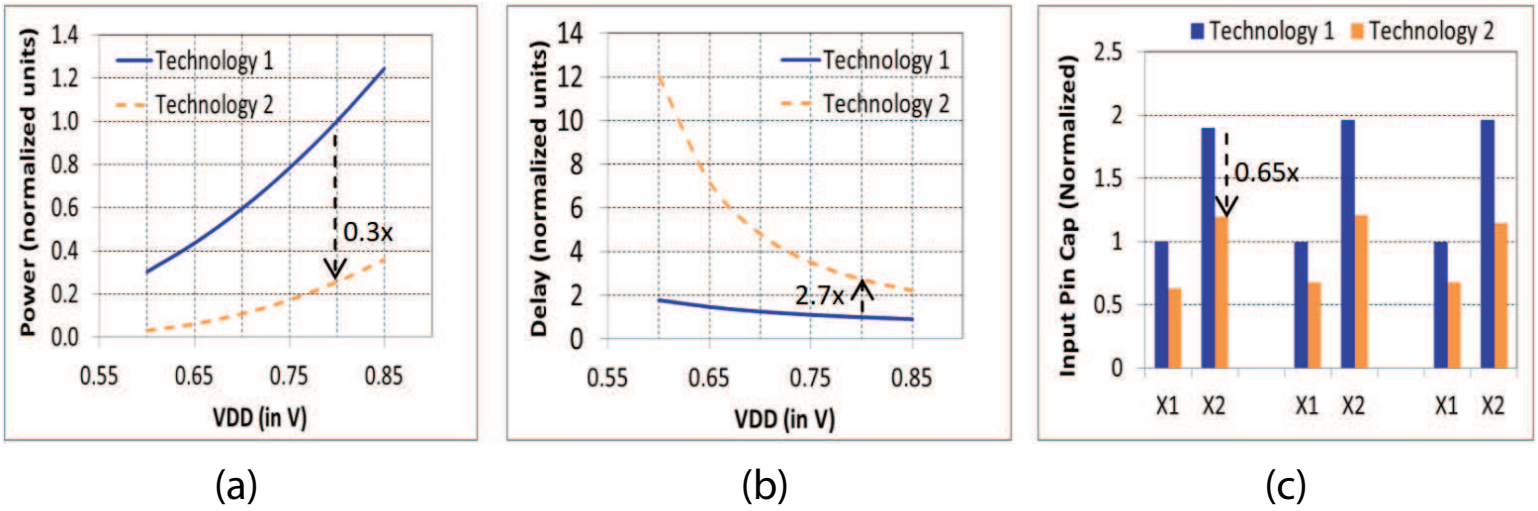


Figure 43: Technology comparison (a) power of inverter chain (b) stage delay in inverter chain (c) pin capacitance for different cells and drivability

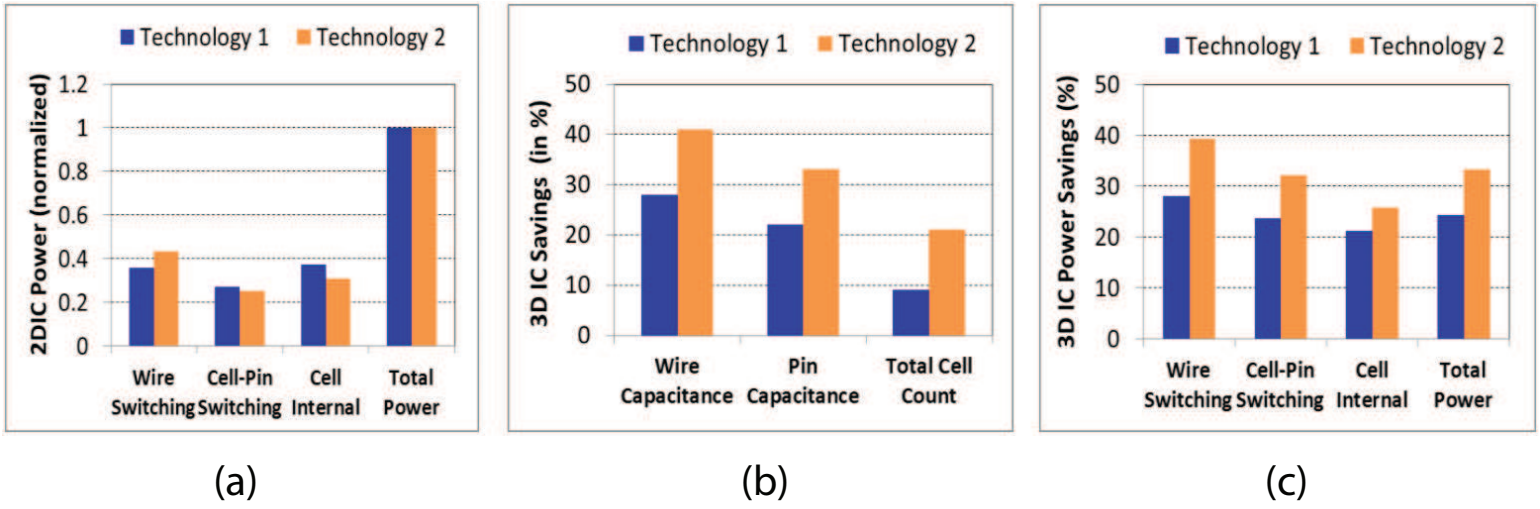


Figure 44: LDPC design results (a) contribution of power components in 2D IC (b) relative 3D IC savings in capacitance and cells (c) relative 3D IC power savings

normalized w.r.t. the total 2D IC power in each technology. Figure 43a shows the contribution of the various components to the total power. As discussed earlier, Technology 1 has higher cell-internal power and higher cell pin capacitance. Therefore, cell internal power has highest contribution and switching of cell pins also add significantly to total power. It is to be noted that LDPC is interconnect dominated circuit. Nevertheless, cell-internal power still has maximum contribution in Technology 1. In contrast, Technology 2 has lower cell power and lower pin capacitance of cells. This not only increases the portion of wire power in total power, but also needs more timing buffers to satisfy timing constraints. The number of additional buffers required is higher in Technology 2, because the cells have relatively lower drive strength compared to Technology 1. As a consequence, the impact of reduction in interconnect length is expected to have higher impact in Technology 2 compared to that in Technology 1.

Figure 44b-c show the detailed comparison of savings obtained in M3D with the two different technologies in terms of capacitance and power. Firstly, Technology 1 has stronger cells and therefore, impact of interconnect on number of timing buffers and total power is less than that of Technology 2. The reduction of interconnect in M3D IC also has lesser impact in Technology 1 compared to Technology 2. Secondly, the input pin capacitance of cells in Technology 1 is much higher than that of Technology 2. Therefore, functional logic without any buffers has much more power contribution in Technology 1. As a result, the buffer and cell size savings in Technology 1 are less than that of Technology 2. Relative wire length reduction due to footprint shrink is similar for both technologies. However, the reduction of higher number of cells in M3D IC in Technology 2 results in further reduction of number of nets and wire length. This results in additional wire-switching power savings in Technology 2 (39% in Technology 2 vs. 28% in Technology 1).

5.3 Summary

The comparison of monolithic 3D ICs and TSV-based 3D ICs is carried out using a foundry 14nm FinFET PDK to find that fine-grained partitioning is not practical with TSVs due to huge size of 3D vias compared to logic gates. Monolithic 3D IC technology, on the other hand, provides true 3D IC benefits in the vertical direction. Three design options were analyzed for 3D IC implementation. It is shown that under best conditions, monolithic 3D can deliver one node PPC advantage, whereas TSV-based 3D designs can achieve only half a node benefit.

The power savings in monolithic 3D ICs using two different foundry technologies is also compared. Monolithic 3D ICs offer significant power savings in both the technologies but the benefits are higher in the technology with lower cell power contribution and smaller input pin cell capacitance.

CHAPTER VI

ADDITIONAL TOPICS

The following project involving 3D ICs and low power device technology impact was also completed along with major work on monolithic 3D ICs

6.1 Near-Threshold Voltage 3D IC Design Study

Near Threshold Computing (NTC) and 3D ICs provide mutual benefits to each other. While NTC designs have an order of magnitude lower power resulting in reduced thermal problems and power delivery demand, 3D ICs help in improving the performance both at the physical design and architecture levels.

6.1.1 Motivation and background

With reduced power dissipation and maximum energy efficiency, near-threshold computing creates a feasible opportunity to successfully tap the advantages of device scaling by utilizing all transistors simultaneously without worrying about thermal issues [21, 15, 13, 33]. However, excessive performance degradation is a major bottleneck. Most of the proposed techniques to improve performance for NTC designs are limited to architectural changes to implement NTC-based parallelism which achieve the desired performance while remaining more energy efficient than its single nominal counter part [57, 20]. Device optimization for lower voltage operation and newer device technologies like fully depleted silicon-on-insulator (FDSOI) with very low leakage are other explored options [38, 18, 6].

3D ICs offer reduced interconnects, reduced footprint, on-chip memory to logic connections, and shorter paths which reduce power and provide potential increase in performance. A memory bandwidth of 63.8GB/s is demonstrated in two tier 3D IC in 130nm process [35]. The first near-threshold 3D IC system was designed with 0.65V for logic and around 0.87V

Table 28: Summary of the three different implementations of OpenSPARC T2 single-core. The number in brackets denote the percentage of total cell count to the nearest integer.

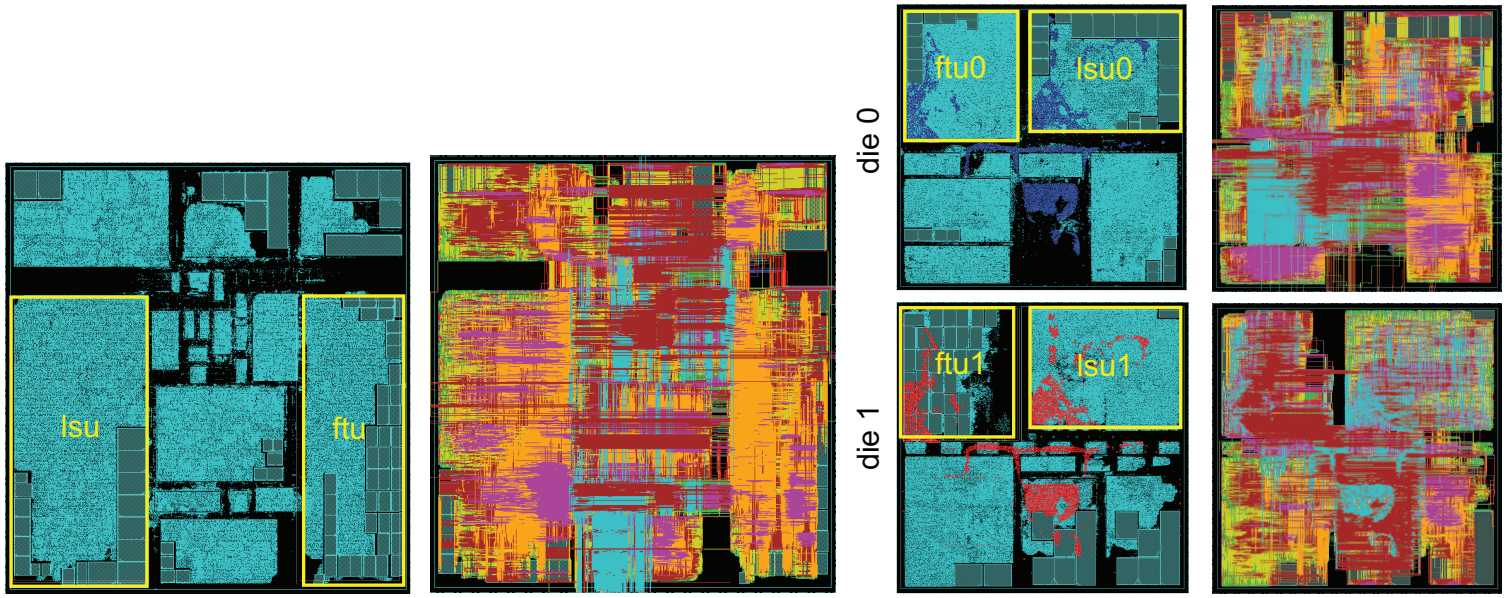
	Nominal 2D IC	NTC 2D IC	NTC 3D IC
Footprint (mm^2)	1.64 x 1.75	1.64 x 1.75	1.20 x 1.20
Max Frequency (MHz)	813.0	**116.3**	**150.6**
Cell Count (x1000)	365.7	366.8	386.4
Buffer Count (x1000)	53.9 (15%)	54.7 (15%)	64.5 (17%)
HVT Cells (x1000)	278.6	253.5	257.6
RVT Cells (x1000)	71.7 (20%)	103.7 (28%)	102.9 (27%)
LVT Cells (x1000)	15.4 (4%)	9.6 (3%)	25.9 (6%)
Wirelength (m)	14.8	14.6	14.7

for SRAM [24]. The authors try to highlight the feasibility of thermal-constrained 3D IC designs by combining it with near-threshold architecture which also results in high energy efficiency. They explore the benefits of cluster-based NTC architecture with 3D stacking in Centip3De and show four-core cluster systems to be 27% more energy efficient while providing 55% more throughput than a one-core cluster system. The cores and caches are in different layers in this 3D implementation. 3D ICs also provide the option of logic on logic folding where logic cells are placed in two or more tiers thereby reducing the signal wirelength [31, 30]. Various techniques such as 3D floorplanning, block folding, metal layer usage control and multi-Vth designs are used. These design techniques are important for low voltage designs. They not only result in lower interconnect switching power but also reduce the timing optimization effort due to shorter paths for the same timing target.

6.1.2 Design and results

6.1.2.1 Design implementation

Full RTL to GDSII block-level implementation of an OpenSPARC T2 single core design in 28nm technology with multi-V_{TH} library is used for this case-study. Two-tier Through Silicon Vias (TSV) based 3D IC at 0.6V is designed and compared with 2D IC at nominal (1.05V) and near-threshold (0.6V) voltages. All the designs are pushed till maximum achievable frequency of operation with no timing violation on any path.



(a) NTC (0.6V) OpenSPARC T2: 2D IC layouts

(b) NTC (0.6V) OpenSPARC T2: 3D IC layouts

Figure 45: Near-V_{TH} (Vdd = 0.6V) OpenSPARC T2 single-core layouts. (a) 2D implementation (footprint 1.75x1.64mm), (b) 3D implementation (footprint = 1.2x1.2mm). Folded blocks (lsu and ftu) are highlighted in yellow. There are 3381 TSVs shown in blue in die0 and the corresponding landing pads are in red in die1 in the placement view. Top-level, lsu, and ifu_ftu have 1531, 1132, and 718 TSVs respectively. All layouts are shown to scale.

Table 29: Power-performance comparison. Numbers in brackets denote percentage relative to nominal 2D.

	Nominal 2D	Near-V_{TH} 2D	Near-V_{TH} 3D
Frequency (MHz)	813.0	116.3 (14%)	150.6 (19%)
Switching Power (mW)	224.3	9.7 (4%)	9.9 (4%)
Internal Power (mW)	633.2	23.9 (4%)	31.6 (5%)
Leakage Power (mW)	16.7	1.2 (7%)	1.4 (8%)
Total Power (mW)	874.2	34.8 (4%)	42.9 (5%)
Power Delay Product (pJ)	1075.3	299.3 (28%)	284.9 (27%)

While multi-V_{TH} optimization helps in improving speed in 2D OpenSPARC T2, the presence of long nets affects the overall timing and also increases power due to increased wirelength. 3D implementation facilitates shortening of nets in general. To reduce the net lengths further, a two stage design folding strategy [31] is implemented. First, the most power hungry blocks in the design are folded followed by two-tier 3D floorplaning. The folding is carried out based on the intra-block architecture such that the highly connected sub-modules remain in the same tier. Based on this folded netlist of the blocks, top level partitioning and 3D floorplanning is carried out to reduce the intra-block wirelength. The folded blocks are kept at the same location in both dies (Figure 45). Using the 3D folding results and die location of the blocks, the netlist connectivity in each die is used to partition the pins of the folded blocks (lsu and ifu_ftu) into the two separate dies. Another important design feature is the intentional use of large white space between blocks in die0 to facilitate optimized TSV insertion and ensure short connections between blocks. However, in the process of allocating white space, the overall silicon area is maintained to be the same in 2D and 3D implementations (Table 28).

6.1.2.2 3D IC performance impact

All the designs are targeted to achieve maximum attainable frequency. It is observed that nominal 2D IC reaches up to 813 MHz (1.23ns clock) while the best frequency of NTC 2D IC is 116.3 MHz (8.6ns clock). 2-tier NTC 3D IC on the other hand beats its 2D counterpart by a significant margin of 29.5% by going up to a frequency of 150.6 MHz (6.64ns clock)

(Table 28). 3D IC has more cells compared to its 2D counterpart at 0.6V. This is because it is possible to insert more buffers in the 3D design to achieve faster clock periods without extra power overhead as the nets are quite short. Short nets result in shorter transition times and lower switching power per net. On the other hand, 2D design has long nets which cannot be optimized even with increased buffer count. The optimization tool modifies the netlist during pre-CTS optimization based on timing and power constraints. Buffers are added, and the type and count of cells change, e.g., a multi-input AND is replaced with multiple 2-input ANDs. Timing is successfully closed for 3D IC at a faster clock compared to 2D IC and there are more such netlist changes for 3D IC. Therefore, 3D IC designs contain more cells apart from extra buffers.

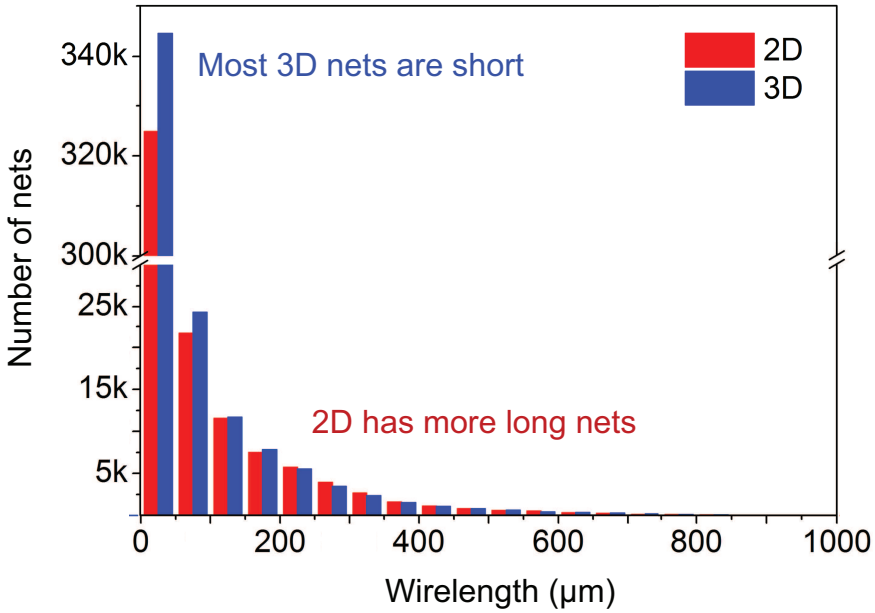


Figure 46: Number of nets in different wirelength bins for NTC implementation with 2D and 3D.

Table 29 shows the results of post-layout power and timing analysis. 3D IC design has more cells due to tighter clock constraints, more low-Vth cells, and run 29.5% faster. Therefore, internal power is higher. However, 3D inter-cell net-switching power is still similar to 2D because of shorter nets, i.e. lower capacitive load. There are 405.6K nets in 3D and 383.6K nets in 2D design at 0.6V but the overall wirelength is almost equal which

implies that the average net length is shorter in 3D IC (Figure 46). More LVT cells in 3D IC result in higher leakage but helps in getting the performance boost. The scaling of voltage in 2D IC domain reduces power by 25X and performance by 7X, resulting in power-delay product (PDP) savings of 3.6X. NTC 3D IC not only increases performance by 29.5% over NTC 2D IC, but also reduces PDP by another 5%.

6.2 Summary

In this chapter, NTC 3D IC design study was presented. Unlike prior works discussing results with single devices and ring-oscillator chains, this study is based on commercial quality full-chip GDSII layouts containing hundreds of thousand of logic gates. The full-chip analysis is based on state-of-the-art RTL-GDSII design flow.

CHAPTER VII

CONCLUSIONS AND FUTURE DIRECTIONS

Monolithic 3D ICs is an emerging technology with the potential to offer significant power performance benefits and continuation of Moore's law. However, as a new technology with innovative fabrication procedure and high integration density, it has inherent issues which have impact on system design. In this dissertation, the design challenges in monolithic 3D ICs have been presented and new CAD solutions have been developed to address them with minimum design overhead. In addition, a foundry based power, performance and cost analysis is also presented to study the place of monolithic 3D IC among other technology options.

In general, the following are the key challenges in practical monolithic 3D IC design: (1) Thermal optimization due to vertical overlap of devices, (2) Power delivery optimization due to increased current demand per unit area and competition of resources between signal and power wires, (3) Handling low performance transistors in the top-tier due to low thermal budget during fabrication, (4) BEOL impact due to use of tungsten and non-negligible 3D net-length and their optimization, and (5) An understanding of impact of transistor technology and comparison with existing technologies in terms of cost effectiveness.

New design methodologies or improved fabrication methods are required to address these challenges. With the objective of understanding the impact of these challenges at the system level and handling them through CAD solutions, the following projects have been presented in this dissertation.

- Fast accurate thermal modeling and optimization for monolithic 3D ICs.
- Full chip impact and optimization of power delivery network in monolithic 3D IC.

- CAD methodology for handling low performance transistors in top-tier of monolithic 3D ICs.
- Tier partitioning strategies to reduce BEOL degradation impact in monolithic 3D ICs.
- Overall Power, performance and cost comparison with 2D IC and TSV-based 3D ICs while understanding the impact of transistor technology on power savings in monolithic 3D ICs.

Near-threshold voltage 3D ICs design study has also been presented in this dissertation as an additional topic targeted for low power designs.

For the objective of thermal optimization, first the unique thermal properties of monolithic 3D ICs were studied and compared with TSV-based 3D ICs. These properties were utilized to develop a methodology to obtain package-aware fast and accurate thermal analysis model for monolithic 3-D ICs with different number of stacking layers using nonlinear regression. These models were verified against full chip FEA thermal simulations. The models were then used in a thermal aware floorplanner to show significant temperature reduction with minimum or no area overhead for both conventional packages with heat sink and mobile packages.

For power delivery optimization in monolithic 3D ICs, the full chip impact of power delivery network has been presented with comparison to the impact in simple 2D designs. The issue becomes much more serious at advanced technology nodes. The role of PDN in the full chip thermal behavior has also been analyzed. Simple yet efficient PDN design styles for wirelength and power reduction were analyzed.

For fabrication related design challenges, the full-chip impact of slower transistors in the top-tier and tungsten in the bottom tier of monolithic 3D ICs were independently assessed. A Tier-Aware M3D (TA-M3D) design flow to handle such slow transistors in the top-tier using industry-quality tools. The critical issues of BEOL impact on the performance of gate-level monolithic 3D ICs were also addressed. A path-based tier partitioning

algorithm was developed to handle the impact of increased resistance of the bottom-tier interconnects with negligible design overhead. The net-based tier partitioning algorithm helps in creating wirelength skew without area skew. This helps in reducing the number of metal layers in bottom-tier without any routing congestion, therefore enabling cost reduction.

Lastly, the impact of transistor technology on power savings in monolithic 3D ICs has been presented. Foundry data based power, performance and cost comparison of monolithic 3D ICs with TSV-based 3D ICs and 2D ICs has been shown.

The research presented in this dissertation is targeted towards addressing many of the design challenges in monolithic 3D ICs with CAD. However, there is more work necessary in both the fabrication, architecture design and physical design aspects. Low temperature device fabrication without compromise on quality is important. Architectural innovation is also essential to exploit the major advantages offered by monolithic 3D ICs in terms of high integration density, lower power and improved performance. Artificial neural network implementation in monolithic 3D IC is one such application. Further CAD development for real 3D place and route tools, transistor-level monolithic 3D IC optimization and memory design are other potential research directions. Apart from all these, near-threshold operation in 3D ICs with improved performance is also interesting. This has been presented as an additional topic in this work.

Therefore, the work in this dissertation will serve as a good starting reference for further development and optimization in addressing key challenges in monolithic 3D ICs to bring it to mass-scale production in the near future.

REFERENCES

[1] ABABEI, C., NAVARATNASOTHIE, S., BAZARGAN, K., and KARYPIS, G., "Multi-objective Circuit Partitioning for Cutsize and Path-Based Delay Minimization," in *ICCAD*, pp. 181–185, Nov 2002.

[2] ATHIKULWONGSE, K., EKPANYAPONG, M., and LIM, S. K., "Exploiting Die-to-Die Thermal Coupling in 3-D IC Placement," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,*, vol. 22, pp. 2145–2155, Oct 2014.

[3] BATUDE, P., SKLENARD, B., FENOUILLET-BERANGER, C., PREVITALI, B., TABONE, C., ROZEAU, O., BILLOINT, O., TURKYILMAZ, O., SARHAN, H., THURIES, S., CIBRARIO, G., BRUNET, L., DEPRAT, F., MICHALLET, J.-E., CLERMIDY, F., and VINET, M., "3D Sequential Integration Opportunities and Technology Optimization," in *IITC/AMC*, pp. 373–376, May 2014.

[4] BATUDE, P., SKLENARD, B., XU, C., PREVITALI, B., SALVO, B. D., and VINET, M., "Low temperature FDSOI devices, a key enabling technology for 3D sequential integration," in *VLSI-TSA*, pp. 1–4, April 2013.

[5] BATUDE, *et al*, P., "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 4, pp. 714–722, 2012.

[6] BEIGNE, E., VALENTIAN, A., GIRAUD, B., THOMAS, O., BENOIST, T., THONNART, Y., BERNARD, S., MORITZ, G., BILLOINT, O., MANEGLIA, Y., FLATRESSE, P., NOEL, J. P., ABOUZEID, F., PELLOUX-PRAYER, B., GROVER, A., CLERC, S., ROCHE, P., COZ, J. L., ENGELS, S., and WILSON, R., "Ultra-Wide Voltage Range designs in Fully-Depleted Silicon-On-Insulator FETs," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2013*, pp. 613–618, March 2013.

[7] BENEVENTI, F., BARTOLINI, A., TILLI, A., and BENINI, L., "An Effective Gray-Box Identification Procedure for Multicore Thermal Modeling," *Computers, IEEE Transactions on*, vol. 63, pp. 1097–1110, May 2014.

[8] BENEVENTI, F., BARTOLINI, A., VIVET, P., DUTOIT, D., and BENINI, L., "Thermal analysis and model identification techniques for a logic + WIDEIO stacked DRAM test chip," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pp. 1–4, March 2014.

[9] BILLOINT, O., SARHAN, H., RAYANE, I., VINET, M., BATUDE, P., FENOUILLET-BERANGER, C., ROZEAU, O., CIBRARIO, G., DEPRAT, F., FUSTIER, A., MICHALLET, J.-E., FAYNOT, O., TURKYILMAZ, O., CHRISTMANN, J.-F., THURIES, S.,

and CLERMIDY, F., "A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool," in *DATE*, pp. 1192–1196, March 2015.

[10] CEYHAN, A., JUNG, M., PANTH, S., LIM, S. K., and NAEEMI, A., "Impact of size effects in local interconnects for future technology nodes: A study based on full-chip layouts," in *IEEE International Interconnect Technology Conference*, pp. 345–348, May 2014.

[11] CEYHAN, A., JUNG, M., PANTH, S., LIM, S. K., and NAEEMI, A., "Evaluating Chip-Level Impact of Cu/Low-k Performance Degradation on Circuit Performance at Future Technology Nodes," *IEEE Transactions on Electron Devices*, vol. 62, pp. 940–946, March 2015.

[12] CHAN, W.-T. J., NATH, S., KAHNG, A. B., DU, Y., and SAMADI, K., "3DIC Benefit Estimation and Implementation Guidance from 2DIC Implementation," in *DAC*, pp. 30:1–30:6, June 2015.

[13] CHANDRAKASAN, A., DALY, D., FINCHELSTEIN, D., KWONG, J., RAMADASS, Y., SINANGIL, M., SZE, V., and VERMA, N., "Technologies for ultradynamic voltage scaling," *Proceedings of the IEEE*, vol. 98, pp. 191–214, Feb 2010.

[14] CHANG, K., ACHARYA, K., SINHA, S., CLINE, B., YERIC, G., and LIM, S. K., "Power benefit study of monolithic 3D IC at the 7nm technology node," in *ISLPED*, pp. 201–206, July 2015.

[15] CHANG, L., FRANK, D., MONTOYE, R., KOESTER, S., JI, B., COTEUS, P., DENNARD, R., and HAENSCH, W., "Practical strategies for power-efficient computing technologies," *Proc. IEEE*, vol. 98, pp. 215–236, Feb 2010.

[16] CHEN, Y., KURSUN, E., MOTSCHMAN, D., JOHNSON, C., and XIE, Y., "Analysis and mitigation of lateral thermal blockage effect of through-silicon-via in 3D IC designs," in *Proc. International Symposium on Low Power Electronics and Design*, pp. 397–402, Aug 2011.

[17] CONG, J., WEI, J., and ZHANG, Y., "A thermal-driven floorplanning algorithm for 3D ICs," in *Proc. IEEE/ACM International Conference on Computer Aided Design.*, pp. 306–313, Nov 2004.

[18] CORSONELLO, P., PERRI, S., and FRUSTACI, F., "Exploring well configurations for voltage level converter design in 28 nm UTBB FDSOI technology," in *Computer Design (ICCD), 2015 33rd IEEE International Conference on*, pp. 499–504, Oct 2015.

[19] DONG, X., ZHAO, J., and XIE, Y., "Fabrication Cost Analysis and Cost-Aware Design Space Exploration for 3-D ICs," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, pp. 1959–1972, Dec 2010.

[20] DRESLINSKI, R. G., ZHAI, B., MUDGE, T., BLAAUW, D., and SYLVESTER, D., "An Energy Efficient Parallel Architecture Using Near Threshold Operation," in *Parallel Architecture and Compilation Techniques, 2007. PACT 2007. 16th International Conference on*, pp. 175–188, Sept 2007.

[21] DRESLINSKI, R., WIECKOWSKI, M., BLAAUW, D., SYLVESTER, D., and MUDGE, T., "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," *Proc. IEEE*, vol. 98, pp. 253–266, Feb 2010.

[22] EMMA, P., BUYUKTOSUNOGLU, A., HEALY, M., KAILAS, K., PUENTE, V., YU, R., HARTSTEIN, A., BOSE, P., and MORENO, J., "3D stacking of high-performance processors," in *Proc. IEEE International Symposium on High Performance Computer Architecture*, pp. 500–511, Feb 2014.

[23] FALKENSTERN, P., XIE, Y., CHANG, Y.-W., and WANG, Y., "Three-dimensional Integrated Circuits (3D IC) Floorplan and Power/Ground Network Co-synthesis," in *Proc. IEEE/ACM Asia South Pacific Design Automation Conference*, pp. 169–174, 2010.

[24] FICK, *et al*, D., "Centip3De : A 3930 DMIPS/W Configurable Near-Threshold 3D Stacked System with 64 ARM Cortex-M3 Cores," in *ISSCC Dig. Tech. Papers*, pp. 190–191, 2012.

[25] FIDUCCIA, C. M. and MATTHEYSES, R. M., "A Linear-Time Heuristic for Improving Network Partitions," in *DAC*, pp. 175–181, June 1982.

[26] HUANG, W., STAN, M., SKADRON, K., K.SANKARANARAYANAN, GHOSH, S., and VELUSAMY, S., "Compact thermal modeling for temperature-aware design," in *DAC*, pp. 878–883, 2004.

[27] HUNG, W.-L., LINK, G., XIE, Y., VIJAYKRISHNAN, N., and IRWIN, M., "Interconnect and thermal-aware floorplanning for 3D microprocessors," in *Proc. International Symposium on Quality Electronic Design*, pp. 6 pp.–104, March 2006.

[28] ITRS. http://www.itrs2.net/, May 2015.

[29] JUAN, D.-C., GARG, S., and MARCULESCU, D., "Statistical Thermal Evaluation and Mitigation Techniques for 3D Chip-Multiprocessors In the Presence of Process Variations," in *DATE*, pp. 1–6, 2011.

[30] JUNG, M., SONG, T., PENG, Y., and LIM, S. K., "Fine-Grained 3-D IC Partitioning Study With a Multicore Processor," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 5, pp. 1393–1401, Oct 2015.

[31] JUNG, M., SONG, T., WAN, Y., LEE, Y.-J., MOHAPATRA, D., WANG, H., TAYLOR, G., JARIWALA, D., PITCHUMANI, V., MORROW, P., WEBB, C., FISCHER, P., and LIM, S. K., "How to Reduce Power in 3D IC Designs: A Case Study with OpenSPARC T2 Core," in *IEEE Custom Integrated Circuits Conference*, pp. 1–4, 2013.

[32] KAHNG, A. B., LIN, B., and SAMADI, K., "Improved On-Chip Router Analytical Power and Area Modeling," in *ASP-DAC*, pp. 241–246, 2010.

[33] KAUL, H. and OTHERS, "Near-threshold voltage (NTV) Design-Opportunities and challenges," in *Proc. ACM Design Automation Conf.*, pp. 1149–1154, June 2012.

[34] KIM, D. H., ATHIKULWONGSE, K., and LIM, S. K., "A Study of Through-Silicon-Via Impact on the 3D Stacked IC Layout," in *ICCAD 2009*, pp. 674–680, Nov 2009.

[35] KIM, *et al*, D. H., "3D-MAPS: 3D Massively Parallel Processor with Stacked Memory," in *ISSCC Dig. Tech. Papers*, pp. 188–189, 2012.

[36] LEE, Y.-J. and LIM, S. K., "Ultrahigh Density Logic Designs Using Monolithic 3-D Integration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, pp. 1892–1905, Dec 2013.

[37] LEE, Y.-J., MORROW, P., and LIM, S. K., "Ultra High Density Logic Designs Using Transistor-Level Monolithic 3D Integration," in *ICCAD*, pp. 539–546, Nov 2012.

[38] LO, C.-W., MEN, L., BRADY, J., and DI, J., "Asynchronous and synchronous designs for low-power FDSOI CMOS process optimized for subthreshold operation at 0.3V VDD," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2015 IEEE*, pp. 1–3, Oct 2015.

[39] LUO, P.-W., ZHANG, C., CHANG, Y.-T., CHENG, L.-C., LEE, H.-H., SHEU, B.-L., SU, Y.-S., KWAI, D.-M., and SHI, Y., "Benchmarking for Research in Power Delivery Networks of Three-Dimensional Integrated Circuits," in *Proc. ACM International Symposium on Physical Design*, pp. 17–24, 2013.

[40] NANGATE15NM. http://www.nangate.com/?page_id=2328, January 2016.

[41] NAYAK, D. K., BANNA, S., SAMAL, S. K., and LIM, S. K., "Power, performance, and cost comparisons of monolithic 3D ICs and TSV-based 3D ICs," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2015 IEEE*, pp. 1–2, Oct 2015.

[42] OPENCORES. http://opencores.org/, October 2012.

[43] ORACLE OPENSPARC T2. http://www.oracle.com, August 2014.

[44] PANTH, S., SAMADI, K., DU, Y., and LIM, S. K., "Tier-partitioning for power delivery vs cooling tradeoff in 3D VLSI for mobile applications," in *Design Automation Conference (DAC), 2015 52nd ACM/EDAC/IEEE*, pp. 1–6, June 2015.

[45] PANTH, S., SAMADI, K., DU, Y., and LIM, S. K., "Design and CAD Methodologies for Low Power Gate-Level Monolithic 3D ICs," in *Low Power Electronics and Design (ISLPED), 2014 ACM/IEEE International Symposium on*, pp. 171–176, Aug 2014.

[46] PANTH, S., SAMADI, K., DU, Y., and LIM, S. K., "Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations," in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1–6, June 2014.

[47] PANTH, S., SAMADI, K., DU, Y., and LIM, S. K., "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," in *Proc. ACM International Symposium on Physical Design*, 2014.

[48] PREDICTIVE TECHNOLOGY MODEL. http://ptm.asu.edu/, September 2012.

[49] RAJENDRAN, B., SHENOY, R. S., WITTE, D. J., CHOKSHI, N. S., DELEON, R. L., TOMPA, G. S., and PEASE, R. F. W., "Low Thermal Budget Processing for Sequential 3-D IC Fabrication," *IEEE Transactions on Electron Devices*, vol. 54, pp. 707–714, April 2007.

[50] SALFORD SYSTEMS. http://www.salford-systems.com/products/mars, May 2013.

[51] SAPATNEKAR, S., "Addressing thermal and power delivery bottlenecks in 3D circuits," in *Proc. Asia and South Pacific Design Automation Conference.*, pp. 423–428, Jan 2009.

[52] SYNOPSYS STANDARD CELL LIBRARIES. https://www.synopsys.com/, August 2014.

[53] TANG, X., TIAN, R., and WONG, D., "Fast evaluation of sequence pair in block placement by longest common subsequence computation," in *Proc. Design, Automation and Test in Europe.*, pp. 106–111, 2000.

[54] WEI, H., WU, T. F., SEKAR, D., CRONQUIST, B., PEASE, R. F., and MITRA, S., "Cooling Three-Dimensional Integrated Circuits using Power Delivery Networks," in *International Electron Devices Meeting*, pp. 14.2.1–14.2.4, 2012.

[55] WONG, S. S. and GAMAL, A. E., "The Prospect of 3D-IC," in *2009 IEEE Custom Integrated Circuits Conference*, pp. 445–448, Sept 2009.

[56] YANG, K., KIM, D. H., and LIM, S. K., "Design quality tradeoff studies for 3D ICs built with nano-scale TSVs and devices," in *Proc. Int. Symp. on Quality Electronic Design*, pp. 740–746, March 2012.

[57] ZHAI, B., DRESLINSKI, R. G., BLAAUW, D., MUDGE, T., and SYLVESTER, D., "Energy efficient near-threshold chip multi-processing," in *Low Power Electronics and Design (ISLPED), 2007 ACM/IEEE International Symposium on*, pp. 32–37, Aug 2007.

[58] ZHOU, P., MA, Y., LI, Z., DICK, R., SHANG, L., ZHOU, H., HONG, X., and ZHOU, Q., "3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits," in *Proc. IEEE/ACM International Conference on Computer-Aided Design.*, pp. 590–597, Nov 2007.

PUBLICATIONS

This dissertation is based on and/or related to the works and results presented in the following publications in print:

[1] **Sandeep Kumar Samal**, Yarui Peng, Yang Zhang, and Sung Kyu Lim, "Design and Analysis of Ultra Low Power Processors Using Sub/Near-Threshold 3D Stacked ICs," International Symposium on Low Power Electronics and Design, 2013.

[2] **Sandeep Kumar Samal**, Kiyoung Kim, Youngchan Kim, Taesung Kim, Hyuk-Jae Lee, Taewhan Kim, and Sung Kyu Lim, "Ultra low power 2-tier 3D stacked subthreshold H.264 intra frame encoder", in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, pp. 1–2, 2013.

[3] **Sandeep Kumar Samal**, Shreepad Panth, Kambiz Samadi, Mehdi Saeidi, Yang Du, and Sung Kyu Lim, "Fast and Accurate Thermal Modeling and Optimization for Monolithic 3D ICs", in *ACM Design Automation Conference*, pp. 1–6, 2014.

[4] **Sandeep Kumar Samal**, Yarui Peng, and Sung Kyu Lim, "Design and Analysis of Ultra Low Power Processors Using Sub/Near-Threshold 3D Stacked ICs," SRC TECHCON Conference, 2014.

[5] Shreepad Panth, **Sandeep Kumar Samal**, Yun Seop Yu, and Sung Kyu Lim, "Design Challenges and Solutions for Ultra-High-Density Monolithic 3D ICs", in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, pp. 1–2, 2014.

[6] Shreepad Panth, **Sandeep Kumar Samal**, Yun Seop Yu, and Sung Kyu Lim, "Design Challenges and Solutions for Ultra-High-Density Monolithic 3D ICs", in *Journal of*

Information and Communication Convergence Engineering, Vol. 12, No. 3, pp. 186–192, 2014.

[7] **Sandeep Kumar Samal**, Kambiz Samadi, Pratyush Kamal, Yang Du, and Sung Kyu Lim, "Full chip impact study of power delivery network designs in monolithic 3D ICs", in *IEEE/ACM International Conference on Computer-Aided Design*, pp. 565–572, 2014.

[8] **Sandeep Kumar Samal**, and Sung Kyu Lim, "Ultralow Power Processor Design with 3D IC Operating at Sub/Near-Threshold Voltages", in *CISS: Nano Devices and Circuit Techniques for Low-Energy Applications and Energy Harvesting*, edited by Chong-Min Kyung, Springer, 2015 (ISBN:978-94-017-9989-8).

[9] **Sandeep Kumar Samal**, Yarui Peng, Mohit Pathak, and Sung Kyu Lim, "Ultra-Low Power Circuit Design with Sub/Near-Threshold 3D IC Technologies," IEEE Transactions on Components, Packaging, and Manufacturing Technology, vol.5, no.7, pp.980-990, July 2015

[10] **Sandeep Kumar Samal**, Yang Li, GuoQing Chen, and Sung Kyu Lim, "Improving performance in near-threshold circuits using 3D IC technology", in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, pp. 1–2, 2015.

[11] Deepak Nayak, Srinivasa Banna, **Sandeep Kumar Samal**, and Sung Kyu Lim, "Power, performance, and cost comparisons of monolithic 3D ICs and TSV-based 3D ICs", in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, pp. 1–2, 2015.

[12] **Sandeep Kumar Samal**, Deepak Nayak, Motoi Ichihashi, Srinivasa Banna, and Sung Kyu Lim, "Impact of transistor technology on power savings in monolithic 3D ICs", in *International Symposium on VLSI Technology, Systems and Application*, pp. 1–2, 2016.

[13] **Sandeep Kumar Samal**, Deepak Nayak, Motoi Ichihashi, Srinivasa Banna, and Sung Kyu Lim, "How to Cope with Slow Transistors in the Top-tier of Monolithic 3D ICs: Design Studies and CAD Solutions", in *IEEE/ACM International Symposium on Low Power Electronics and Design*, pp. 320–325, 2016.

[14] **Sandeep Kumar Samal**, Deepak Nayak, Motoi Ichihashi, Srinivasa Banna, and Sung Kyu Lim, "Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology", in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, pp. 1–2, Oct 2016.

[15] **Sandeep Kumar Samal**, Shreepad Panth, Kambiz Samadi, Mehdi Saedi, Yang Du, and Sung Kyu Lim, "Adaptive Regression-Based Thermal Modeling and Optimization for Monolithic 3-D ICs", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, pp. 1707–1720, Oct 2016.

[16] **Sandeep Kumar Samal**, Deepak Nayak, Motoi Ichihashi, Srinivasa Banna, and Sung Kyu Lim, "Tier partitioning strategy to mitigate BEOL degradation and cost issues in monolithic 3D ICs", in *IEEE/ACM International Conference on Computer-Aided Design*, pp. 1–7, 2016.

[17] Jiajun Shi, Deepak Nayak, Srinivasa Banna, Robert Fox, Srikanth Samavedam, **Sandeep Kumar Samal**, and Sung Kyu Lim, "A 14nm Finfet Transistor-Level 3D Partitioning Design to Enable High-Performance and Low-Cost Monolithic 3D IC", in *IEEE International Electron Devices Meeting*, 2016.

[18] **Sandeep Kumar Samal**, Kambiz Samadi, Pratyush Kamal, Yang Du, and Sung Kyu Lim, "Full Chip Impact Study of Power Delivery Network Designs in Gate-Level Monolithic 3D ICs", in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2016 (doi:10.1109/TCAD.2016.2616377).

[19] **Sandeep Kumar Samal**, GuoQing Chen, and Sung Kyu Lim, "Improving Performance Under Process and Voltage Variations in Near-Threshold Computing Using 3D ICs", in *ACM Journal of Emerging Technologies in Computing*, to appear

[20] Shreepad Panth, **Sandeep Kumar Samal**, Kambiz Samadi, Yang Du, and Sung Kyu Lim, "Tier Degradation of Monolithic 3D ICs: A Power Performance Study at Different Technology Nodes", in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, to appear.

In addition, the author has completed works unrelated to this dissertation presented in the following publications in print:

[1] Tianchen Wang, **Sandeep Kumar Samal**, Sung Kyu Lim, and Yiyu Shi, "A novel entropy production based full-chip TSV fatigue analysis", in *IEEE/ACM International Conference on Computer-Aided Design*, pp. 744–751, 2015.

[2] **Sandeep Kumar Samal**, GuoQing Chen, and Sung Kyu Lim, "Machine Learning Based Variation Modeling and Optimization for 3D ICs", in *Journal of Information and Communication Convergence Engineering*, Vol. 14, No. 4, pp. 258-267, 2016.

[3] Tiantao Lu, Caleb Serafy, Zhiyuan Yang, **Sandeep Kumar Samal**, Sung Kyu Lim, and Ankur Srivastava, "TSV-based 3D ICs: Design Methods and Tools", in *IEEE Transactions on Compute-Aided Design of Integrated Circuits and Systems*, to appear.

VITA

Sandeep Kumar Samal was born in Sahaspur, Odisha, India, in 1990. He received his B.Tech (Hons.) in Electronics and Electrical Communication Engineering from Indian Institute of Technology Kharagpur in 2012. He also received M.S. in Electrical and Computer Engineering from Georgia Institute of Technology, Atlanta, USA in 2013, where he is currently a PhD candidate.

From 2012 to present, he has been a graduate research assistant in Georgia Tech Computer Aided Design (GTCAD) laboratory led by Dr. Sung Kyu Lim. His primary research is in the area of 3D IC design, modeling and analysis with major focus on monolithic 3D ICs. His other research interests include low power and reliable digital design and advanced technology-design co-optimization.